KernSmoothIRT: An R Package for Kernel Smoothing in Item Response Theory

Angelo Mazza
University of Catania

Antonio Punzo
University of Catania

Brian McGuire
Montana State University

Abstract

Item response theory (IRT) models are a class of statistical models used to describe the response behaviors of individuals to a set of items. They are adopted by researchers in social science, particularly in the analysis of performance or attitudinal data, in psychology, education, medicine, marketing and other fields where testing is relevant. Although most IRT analyses use parametric models, they rely on assumptions that often are not satisfied. In such cases, a nonparametric approach might be preferable; nevertheless, there are not many software applications allowing it.

In this paper we present the R package **KernSmoothIRT**. It implements kernel smoothing for the estimation of item and option characteristic curves, as well the production of several test and subject based plots. In order to show the package capabilities, two real datasets are used, one employing multiple choices, and the other scaled responses.

Keywords: kernel smoothing, item response theory, principal component analysis, probability simplex.

1. Introduction

In psychometrics and educational testing the analysis of the relation between latent continuous variables and observed dichotomous/polytomous variables is known as item response theory (IRT). Observed variables arise from a test or a questionnaire composed by several items of one of two types: *multiple-choice items*, in which only one option is designed to be correct, and *rating scale items*, in which a different weight is attributed to each item's option (polytomous weighting). Multiple choice items may be viewed as scale items where one option receives a weight of one and the others a weight of zero (dichotomous weighting). Naturally, a set of items can be a mixture of these two types of items.

Our notation and framework can be summarized as follows. Consider the responses of a n-dimensional set $\mathcal{S} = \{S_1, \dots, S_i, \dots, S_n\}$ of subjects to a k-dimensional sequence $\mathcal{I} = \{I_1, \dots, I_j, \dots, I_k\}$ of items. Let $\mathcal{O}_j = \left\{O_{j1}, \dots, O_{jl}, \dots, O_{jm_j}\right\}$ be the m_j-dimensional set of options conceived for $I_j \in \mathcal{I}$, and let x_{jl} be the weight attributed to O_{jl}. The actual response of S_i to I_j can be so represented as a selection vector $\boldsymbol{y}_{ij} = \left(y_{ij1}, \dots, y_{ijm_j}\right)'$, where $\boldsymbol{y}_{ij}$ is

an observation from the random variable $\boldsymbol{Y}_{ij}$ and $y_{ijl} = 1$ if the option O_{jl} is selected, and 0 otherwise. From now on it will be assumed that, for each item $I_j \in \mathcal{I}$, the subject selects one and only one of the m_j options in $\mathcal{O}_j$; omitted responses are permitted. The central problem in polytomous IRT, with reference to a generic option O_{jl} of I_j, is the specification of a mathematical model describing the probability of selecting O_{jl} as a function of ϑ (the discussion is here restricted to models for items that measure one continuous latent variable, *i.e.*, *unidimensional latent trait models*). According to Ramsay (1991), this function, or curve, will be referred to as option characteristic curve (OCC), and it will be denoted with

$$p_{jl}(\vartheta) = \mathsf{P}(\text{select } O_{jl} \,|\vartheta) = \mathsf{P}(Y_{jl} = 1 \,|\vartheta), \tag{1}$$

$j = 1, \ldots, k$, $l = 1, \ldots, m_j$. For example, in the analysis of multiple-choice items, which hastypically relied on numerical statistics such as the p values (proportion of subjects selecting each option) and the point biserial correlation (quantifying item discrimination), it might be more informative to take into account all of the OCCs (Lei, Dunbar, and Kolen 2004). Moreover, the OCCs are the starting points for a wide range of IRT analyses (see, *e.g.*, Baker and Kim 2004).

With the aim to estimate the OCCs, in analogy with the classic statistical modelling, at least two routes are possible. The first, and most common, is the *parametric* one (PIRT: parametric IRT), in which a parametric structure is assumed so that the estimation of an OCC is reduced to the estimation of a vector parameter, of dimension varying from model to model, for each item in $\mathcal{I}$ (see, *e.g.*, Thissen and Steinberg 1986; van der Linden and Hambleton 1997; Ostini and Nering 2006; Nering and Ostini 2010, to have an idea of the existing PIRT models). This vector is usually considered to be of direct interest and its estimate is often used as a summary statistic to describe some aspects, such as difficulty and discrimination, of the corresponding item I_j (see Lord 1980). The second route is the *nonparametric* one (NIRT: nonparametric IRT), in which estimation is made directly on $\boldsymbol{y}_{ij}$, $i = 1, \ldots, n$ and $j = 1, \ldots, k$, without assuming any mathematical form for the OCCs, in order to obtain more flexible estimates which, according to van der Linden and Hambleton (1997, p. 348), can be assumed to be closer to the true OCCs than those provided by PIRT models. Accordingly, Ramsay (1997) argues that NIRT might become the reference approach unless there are substantive reasons for preferring a certain parametric model. Generally, the main advantage of NIRT models are flexibility and computational convenience. Moreover, although nonparametric models are not characterized by parameters of direct interest, they encourage the graphical display of results; Ramsay (1997, p. 384), by personal experience, confirms the communication advantage of an appropriate display over numerical summaries. These are only some of the motivations which justify the growing in NIRT research in recent years; other considerations can be found in Junker and Sijtsma (2001) who identify three broad motivations for the development and continued interest in NIRT.

Among the NIRT models, kernel smoothing (Ramsay 1991) is a promising option, due to conceptual simplicity and practical and theoretical properties. The computer software TestGraf (Ramsay 2000) performs kernel smoothing estimation of OCCs and allows for other related graphical analyses based on them. In this paper we present the R (R Development Core Team 2011) package **KernSmoothIRT**, available from CRAN (`http://CRAN.R-project.org/`), which offers most of the TestGraf features and adds some related functionalities. Note that, although R is well-provided with PIRT techniques (see de Leeuw and Mair 2007 and Wickelmaier, Strobl, and Zeileis 2012), it does not offer nonparametric analyses, of the type

described above, in IRT. Nonparametric smoothing techniques of the kind found in **KernSmoothIRT** are commonly used and often cited exploratory statistical tools; as evidence, consider the number of times in which classical statistical studies use the functions `density` and `ksmooth`, both in the **stats** package, for kernel smoothing estimation of a density or regression function.

The paper is organized as follows. Section 2 discusses the problem of estimating abilities in the nonparametric context. Then, starting from Ramsay (1991), Section 3 retraces kernel smoothing estimation of the OCCs and Section 4 illustrates other useful IRT functions based on these estimates. The relevance of the package is shown, via two real data sets, in Section 5, and conclusions are finally given in Section 6.

2. Estimating abilities

Consider any strictly monotonic transformation $\tau = g(\vartheta)$ of the ability continuum. Then

$$p_{jl}(\vartheta) = p_{jl}\left\{g^{-1}\left[g(\vartheta)\right]\right\} = p_{jl}\left[g^{-1}(\tau)\right] = p^*_{jl}(\tau), \tag{2}$$

where the function $p^*_{jl} = p_{jl} \circ g^{-1}$ is the equivalent OCC relative to the new ability continuum τ; thus, the choice of scale becomes perfectly arbitrary (Bartholomew 1983). This *lack of identifiability*, expressed more elegantly by Samejima (1981), implies that estimation of the functions $p_{jl}(\vartheta)$, are invariant with respect to monotone transformations of their domain. It is interesting to note that this lack of identifiability is recognized in the marginal maximum likelihood (MML; Bock and Lieberman 1970; Bock and Aitkin 1981) estimation procedures for the item parameters of parametric models, where the choice of the prior density for ϑ is regarded as to some degree arbitrary. Consequently, only rank order considerations make sense for the n ability estimates. Nevertheless, if monotone transformations of the rank ordering belong to a smooth family, and the assumption that probabilities do not change discontinuously over the ability *continuum* is reasonable, then the analysis also yields topological information in the sense that two points positioned close to each other will continue to be close under all "reasonable" transformations.

Let T_i be a statistic associated to each subject's response pattern. The total score

$$T_i = \sum_{j=1}^{k}\sum_{l=1}^{m_j} y_{ijl}x_{jl}$$

is the most obvious choice. As suggested in Ramsay (1991, p. 615) and Ramsay (2000, pp. 25–26), to determine the estimates $\widehat{\vartheta}_i$ starting from the values of T_i, one could:

1. estimate the relative rank r_i of S_i by ranking the values T_i. Operationally, for shorter tests and larger number of subjects, many ties in the values of the statistic T may occur. To minimize possible biases due to the order in which tests results are recorded, **KernSmoothIRT** randomizes the ordering of subjects with the same T. Thus, $r_i = R_i/(n+1)$, where $R_i \in \{1,\ldots,n\}$ represents the position of S_i in the randomized ordering;

2. replace r_i by the quantile $\widehat{\vartheta}_i$ of some distribution function F that is seen to be appropriate. The estimated ability value for S_i so becomes $\widehat{\vartheta}_i = F^{-1}(r_i)$. In these

terms, the denominator $n+1$ of r_i avoids an infinity value for the biggest $\widehat{\vartheta}_i$ when $\lim_{\vartheta \to +\infty} F(\vartheta) = 1^-$.

The choice of F is equivalent to the choice of the ϑ-metric. Historically, the standard Gaussian distribution $F = \Phi$ has been heavily used (see Bartholomew 1988, for general arguments and some evidence supporting this choice); it is also one of the most commonly used in applications of the parametric models, to which the kernel model is often compared. Logically, other continuous distributions are not excluded. For example, users who think of ability as percentages may prefer a distribution on $[0,1]$ such as the Beta – a Beta$(2.5, 2.5)$ looks very much like a standard Gaussian (Ramsay 1991) – or the uniform if the relative ranks r_i have to be directly used. **KernSmoothIRT** permits to the user to specify F by all the classical continuous distributions implemented in R.

Since latent ability estimates are rank-based, they are usually referred to as *ordinal ability estimates*. Note that even a substantial amount of error in the ranks has only a small impact on the estimated curve values. This can be demonstrated both by mathematical analysis and through simulated data (see Ramsay 1991, 2000, and Douglas 1997 for further details).

3. Kernel smoothing of OCCs

Ramsay (1991, 1997) popularized nonparametric estimation of OCCs by proposing nonparametric regression methods, based on kernel smoothing approaches, which are implemented in the TestGraf program (Ramsay 2000). The basic idea of kernel smoothing is to obtain a nonparametric estimate of the OCC by taking a (local) weighted average (Altman 1992; Eubank 1988; Härdle 1990; Härdle 1992; Simonoff 1996) of the form

$$\widehat{p}_{jl}(\vartheta) = \sum_{i=1}^{n} w_{ij}(\vartheta) Y_{ijl}, \tag{3}$$

where the weights $w_{ij}(\vartheta)$ are defined so as to be maximal when $\vartheta = \vartheta_i$ and to be smoothly non-increasing as $|\vartheta - \vartheta_i|$ increases. The need to keep $\widehat{p}_{jl}(\vartheta) \in [0,1]$, for each $\vartheta \in \mathbb{R}$, requires the additional constraints $w_{ij}(\vartheta) \geq 0$ and $\sum_{i=1}^{n} w_{ij}(\vartheta) = 1$; as a consequence, it is preferable to use Nadaraya-Watson weights (Nadaraya 1964; Watson 1964) of the form

$$w_{ij}(\vartheta) = \frac{K\left(\dfrac{\vartheta - \vartheta_i}{h_j}\right)}{\displaystyle\sum_{r=1}^{n} K\left(\dfrac{\vartheta - \vartheta_r}{h_j}\right)}, \tag{4}$$

where $h_j > 0$ is the *smoothing parameter* (also known as *bandwidth*) controlling the amount of smoothness (in terms of bias-variance trade-off), while K is the *kernel function*, a nonnegative, continuous ($\widehat{p}_{jl}$ inherits the continuity from K) and usually symmetric function that is non-increasing as its argument moves further from zero.

Since the performance of (3) largely depends on the choice of h_j, rather than on the kernel function (the theoretical background of this observation can be found, *e.g.*, in Marron and Nolan 1988), a simple Gaussian kernel $K(u) = \exp\left(-u^2/2\right)$ is often preferred (this is the only setting available in TestGraf). Nevertheless, **KernSmoothIRT** allows for other common

choices such as the uniform kernel, $K(u) = \mathbb{I}_{[-1,1]}(u)$, and the quadratic kernel $K(u) = \left(1-u^2\right)\mathbb{I}_{[-1,1]}(u)$, where $\mathbb{I}_A(u)$ represents the indicator function assuming value 1 on A and 0 otherwise. The bandwidth h_j, in contrast to both Ramsay (1991) and TestGraf, may vary from item to item (as highlighted by its subscript). This is an important aspect, since different items included in a test may not require the same amount of smoothing to obtain smooth curves (see Lei *et al.* 2004, p. 8).

Unlike the standard kernel regression estimators, in (3) the dependent variable is a binary variable Y_{jl} and the independent one is the latent variable ϑ. Although ϑ cannot be directly observed, kernel smoothing can still be used, but each ϑ_i in (3) must be replaced with a reasonable estimate $\widehat{\vartheta}_i$ (Ramsay 1991), resulting in an estimate of the form

$$\widehat{p}_{jl}(\vartheta) = \sum_{i=1}^{n} \widehat{w}_i(\vartheta) Y_{ijl}, \tag{5}$$

where

$$\widehat{w}_i(\vartheta) = \frac{K\left(\dfrac{\vartheta - \widehat{\vartheta}_i}{h_j}\right)}{\displaystyle\sum_{r=1}^{n} K\left(\dfrac{\vartheta - \widehat{\vartheta}_r}{h_j}\right)}.$$

As underlined in Ramsay (1991), another thing should be noted. The denominator of equation (5) is in effect (proportional to) a Rosenblatt-Parzen kernel estimator (see, *e.g.*, Silverman 1986) of the ability density function $f(\vartheta)$. Although this density is already known, in the sense of being determined by the choice of the quantile distribution F, and consequently could be replaced by the actual density, this substitution is not recommended because it might result in occasional values of $\widehat{p}_{jl}$ slightly outside of the natural interval $[0,1]$.

Regarding the statistical properties of this method, Douglas (1997) shows, for the dichotomous case, that although any $\widehat{p}_{j1}(\vartheta)$ is an empirical regression estimate of Y_{j1} on a total score transformation, it can consistently estimate the true $p_{j1}(\vartheta)$. The author argues that this asymptotic result can easily be extended to the polytomous case. Moreover, Douglas (2001) proves that, for long tests, there is only one correct IRT model for a given choice of F, and nonparametric methods (including the kernel estimation approach) can consistently estimate it. Thus, following the idea of Douglas and Cohen (2001), if nonparametric estimated curves are meaningfully different from parametric ones, this parametric model – defined on the particular scale determined by F – is an uncorrected model for the data. In order to make this comparison valid, it is fundamental that the same F is used for both nonparametric and parametric curves. For example, if MML (that typically assumes a Gaussian distribution for ϑ) is selected to fit a parametric model, kernel estimates represented on this same distribution $F = \Phi$ can be compared to it. Summarizing, in the choice of a parametric family, visual inspections of the estimated kernel curves can be useful.

3.1. Operational aspects

Operationally, the kernel OCC is evaluated on a finite grid, $\vartheta_1, \dots, \vartheta_s, \dots, \vartheta_q$, of q equally-spaced values spanning the range of the $\widehat{\vartheta}_i$'s, so that the distance between two consecutive points is δ. Thus, starting from the values of Y_{ijl} and $\widehat{\vartheta}_i$, by grouping we can define the two

sequences of q values

$$\widetilde{Y}_{sjl} = \sum_{i=1}^{n} \mathbb{I}_{[\vartheta_s - \delta/2, \vartheta_s + \delta/2)}\left(\widehat{\vartheta}_i\right) Y_{ijl} \quad \text{and} \quad V_s = \sum_{i=1}^{n} \mathbb{I}_{[\vartheta_s - \delta/2, \vartheta_s + \delta/2)}\left(\widehat{\vartheta}_i\right).$$

Up to a scale factor, the sequence $\widetilde{Y}_{sjl}$ is a grouped version of Y_{ijl}, while V_s is the corresponding number of subjects in that group. It follows that

$$\widehat{p}_{jl}(\vartheta) \approx \frac{\sum_{s=1}^{q} K\left(\frac{\vartheta - \vartheta_s}{h_j}\right) \widetilde{Y}_{sjl}}{\sum_{s=1}^{q} K\left(\frac{\vartheta - \vartheta_s}{h_j}\right) V_s}, \qquad \vartheta \in \left\{\vartheta_1, \dots, \vartheta_s, \dots, \vartheta_q\right\}. \tag{6}$$

The denominator remains an estimate of $f(\vartheta)$, except for the same scale factor that multiplies $\widetilde{Y}_{sjl}$.

3.2. Cross-validation selection for the bandwidth

Two of the most frequently used methods of bandwidth selection are the plug-in method and the cross-validation (for a more complete treatment of these methods see, *e.g.*, Härdle 1992).

The former approach, widely diffuse in the context of kernel density estimation, often leads to rules of thumb. In particular, for the Gaussian kernel density estimator, under the assumption of normality for the true but unknown distribution, the common rule of thumb of Silverman (1986, p. 45) may be formulated, in our context, as

$$h = 1.06\sigma_{\vartheta} n^{-1/5}, \tag{7}$$

where σ_{ϑ} – that in the original framework is a sample estimate – simply represents the standard deviation of ϑ, according to the "known" distribution F. Note that, in our context, this way of proceeding leads to the use of the same bandwidth for all the items. In Härdle (1992, p. 187) a conversion table of (7), for the other commonly used kernel functions, can also be found. However, for nonparametric regression, such a choice is not natural; the theory, indeed, shows that the optimal bandwidth depends on the curvature in the conditional mean, regardless of the marginal density – $f(\vartheta)$ in our case – of the regressor(s) for which the rule of thumb is designed. Nevertheless, motivated by the need to have fast automatically generated kernel estimates, this rule represents the default value of the function `ksIRT` of **KernSmoothIRT**; in these terms note that (7), with $\sigma_{\vartheta} = 1$, is the unique approach considered in `TestGraf`.

The second approach, cross-validation, although it requires a considerably higher computational effort, is nevertheless simple to understand and natural for nonparametric regression. Ordinary cross-validation has been widely studied in the setting of nonparametric kernel regression (see, *e.g.*, Rice 1984; Wong 1983). Its description, in our context, is as follows. Let $\boldsymbol{y}_j = \left(\boldsymbol{y}_{1j}, \dots, \boldsymbol{y}_{ij}, \dots, \boldsymbol{y}_{nj}\right)$ be the $m_j \times n$ selection matrix referred to I_j. Moreover, let

$$\widehat{\boldsymbol{p}}_j(\vartheta) = \left(\widehat{p}_{j1}(\vartheta), \dots, \widehat{p}_{jm_j}(\vartheta)\right)'$$

be the m_j-dimensional vector of kernel-estimated probabilities, for I_j, at the evaluation point ϑ. The probability kernel estimator evaluated in ϑ, for I_i, can thus be rewritten in the

following form

$$\widehat{\boldsymbol{p}}_j(\vartheta) = \sum_{i=1}^{n} \widehat{w}_{ij}(\vartheta)\, \boldsymbol{y}_{ij} = \boldsymbol{y}_j \widehat{\boldsymbol{w}}_j(\vartheta)$$

where $\widehat{\boldsymbol{w}}_j(\vartheta) = \left(\widehat{w}_{1j}(\vartheta), \ldots, \widehat{w}_{ij}(\vartheta), \ldots, \widehat{w}_{nj}(\vartheta)\right)'$ denotes the vector of weights.

In detail, cross-validation simultaneously fits and smooths the data contained in $\boldsymbol{y}_j$ by removing one "data point" $\boldsymbol{y}_{ij}$ at a time, estimating the value of $\boldsymbol{p}_j$ at the correspondent ordinal ability estimate $\widehat{\vartheta}_i$, and then comparing the estimate to the omitted, observed value. So the cross-validation statistic, $\mathsf{CV}(h_j)$, is

$$\mathsf{CV}(h_j) = \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{y}_{ij} - \widehat{\boldsymbol{p}}_j^{(-i)}\left(\widehat{\vartheta}_i\right)\right)' \left(\boldsymbol{y}_{ij} - \widehat{\boldsymbol{p}}_j^{(-i)}\left(\widehat{\vartheta}_i\right)\right), \tag{8}$$

where

$$\widehat{\boldsymbol{p}}_j^{(-i)}\left(\widehat{\vartheta}_i\right) = \frac{\displaystyle\sum_{\substack{r=1\\ r\neq i}}^{n} K\left(\frac{\widehat{\vartheta}_i - \widehat{\vartheta}_r}{h_j}\right) \boldsymbol{y}_{rj}}{\displaystyle\sum_{\substack{r=1\\ r\neq i}}^{n} K\left(\frac{\widehat{\vartheta}_i - \widehat{\vartheta}_r}{h_j}\right)}$$

is the estimated vector of probabilities at $\widehat{\vartheta}_i$ computed by removing the observed selection vector $\boldsymbol{y}_{ij}$. The value of h_j that minimizes $\mathsf{CV}(h_j)$ is referred to as the cross-validation smoothing parameter, h_j^{CV}, and it is possible to find it by systematically searching across a suitable smoothing parameter region.

3.3. Pointwise confidence intervals

In visual inspection and graphical interpretation of the estimated kernel curves, pointwise confidence intervals at the evaluation points $\vartheta \in \mathbb{R}$ provide relevant information, because they indicate the extent to which the kernel OCCs are well defined across the range of ϑ considered. Moreover, they are useful when nonparametric and parametric models are compared.

Since $\widehat{p}_{jl}(\vartheta)$ is a linear function of the data, as can be easily seen from (5), and being $Y_{ijl} \sim \text{Ber}\left[p_{jl}\left(\widehat{\vartheta}_i\right)\right]$,

$$\begin{aligned}
\mathsf{VAR}\left[\widehat{p}_{jl}(\vartheta)\right] &= \sum_{i=1}^{n} \left[\widehat{w}_i(\vartheta)\right]^2 \mathsf{VAR}\left(Y_{ijl}\right) \\
&= \sum_{i=1}^{n} \left[\widehat{w}_i(\vartheta)\right]^2 p_{jl}\left(\widehat{\vartheta}_i\right)\left[1 - p_{jl}\left(\widehat{\vartheta}_i\right)\right].
\end{aligned}$$

The above formula holds if independence of the Y_{ijl}s is assumed and possible error variation in the arguments, $\widehat{\vartheta}_i$, are ignored (Ramsay 1991). Substituting p_{jl} for $\widehat{p}_{jl}$ yields the $(1-\alpha)\cdot 100\%$ pointwise confidence intervals

$$\widehat{p}_{jl}(\vartheta) \mp z_{1-\frac{\alpha}{2}} \sqrt{\sum_{i=1}^{n} \left[\widehat{w}_i(\vartheta)\right]^2 \widehat{p}_{jl}\left(\widehat{\vartheta}_i\right)\left[1 - \widehat{p}_{jl}\left(\widehat{\vartheta}_i\right)\right]}, \tag{9}$$

where $z_{1-\frac{\alpha}{2}}$ is such that $\Phi\left[z_{1-\frac{\alpha}{2}}\right] = 1-\frac{\alpha}{2}$.

4. Functions related to the OCCs

Once the kernel estimates of the OCCs are obtained, several other quantities can be computed based on them. In what follows we will give a concise list of the most important ones. In these terms, to facilitate the interpretation of the OCCs, as well as of other output-plots of KernSmoothIRT, it may be preferred to use the expected total score

$$\tau(\vartheta) = \sum_{j=1}^{k}\sum_{l=1}^{m_j} \widehat{p}_{jl}(\vartheta)\, x_{jl}, \tag{10}$$

in substitution of ϑ, as display variable on the x-axis. This possibility is considered in KernSmoothIRT through the option `axistype` of the function `plot.ksIRT`. Note that, although it can happen that (10) fails to be completely increasing in ϑ, this event is rare and tends to affect the plots only at extreme trait levels.

4.1. Item characteristic curve

In analogy with the dichotomous case, and starting from (1), in order to obtain a single function for each item in $\mathcal{I}$ it is possible to define the expected value of the score $X_j = \sum_{l=1}^{m_j} x_{jl} Y_{jl}$, conditional on a given value of ϑ (see, *e.g.*, Chang and Mazzeo 1994), as follows

$$e_j(\vartheta) = \mathsf{E}\left(X_j \,|\vartheta\right) = \sum_{l=1}^{m_j} x_{jl} p_{jl}(\vartheta), \tag{11}$$

$j = 1,\ldots,k$, that takes values in $\left[\min\left\{x_{j1},\ldots,x_{jm_j}\right\}, \max\left\{x_{j1},\ldots,x_{jm_j}\right\}\right]$. The function $e_j(\vartheta)$ is commonly known as item characteristic curve (ICC) and can be viewed (Lord 1980) as a regression of the item score X_j onto the ϑ scale. Naturally, for dichotomous and multiple-choice IRT models, the ICC coincides with the OCC referred to the correct option.

Starting from (11), it is straightforward to define the kernel ICC estimator as follows

$$\widehat{e}_j(\vartheta) = \sum_{l=1}^{m_j} x_{jl}\widehat{p}_{jl}(\vartheta) = \sum_{l=1}^{m_j} x_{jl} \sum_{i=1}^{n} \widehat{w}_{ij}(\vartheta)\, Y_{ijl} = \sum_{i=1}^{n} \widehat{w}_{ij}(\vartheta) \sum_{l=1}^{m_j} x_{jl} Y_{ijl}. \tag{12}$$

For the ICC, in analogy with Section 3.3, the $(1-\alpha)\cdot 100\%$ pointwise confidence interval is given by

$$\widehat{e}_j(\vartheta) \mp z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\mathsf{VAR}\left[\widehat{e}_j(\vartheta)\right]}}, \tag{13}$$

and, since $Y_{ijl}Y_{ijt} \equiv 0$ for $l \neq t$, one has

$$\mathsf{VAR}\left[\widehat{e}_j(\vartheta)\right] = \sum_{i=1}^{n} \left[\widehat{w}_{ij}(\vartheta)\right]^2 \mathsf{VAR}\left(\sum_{l=1}^{m_j} x_{jl} Y_{ijl}\right)$$

where

$$\begin{aligned}\mathsf{VAR}\left(\sum_{l=1}^{m_j} x_{jl}Y_{ijl}\right) &= \sum_{l=1}^{m_j} x_{jl}^2 \mathsf{VAR}\left(Y_{ijl}\right) + \sum_{l=1}^{m_j}\sum_{t\neq l} x_{jl}x_{jt}\mathsf{COV}\left(Y_{ijl},Y_{ijt}\right)\\ &= \sum_{l=1}^{m_j} x_{jl}^2 \mathsf{VAR}\left(Y_{ijl}\right) - \sum_{l=1}^{m_j}\sum_{t\neq l} x_{jl}x_{jt}\mathsf{E}\left(Y_{ijl}\right)\mathsf{E}\left(Y_{ijt}\right)\\ &= \sum_{l=1}^{m_j} x_{jl}^2 p_{jl}\left(\widehat{\vartheta}_i\right)\left[1-p_{jl}\left(\widehat{\vartheta}_i\right)\right] - \sum_{l=1}^{m_j}\sum_{t\neq l} x_{jl}x_{jt}p_{jl}\left(\widehat{\vartheta}_i\right)p_{jt}\left(\widehat{\vartheta}_i\right).\end{aligned}$$

Substituting p_{jl} with $\widehat{p}_{jl}$ in $\mathsf{VAR}\left[\widehat{e}_i\left(\vartheta\right)\right]$, one obtains $\widehat{\mathsf{VAR}\left[\widehat{e}_i\left(\vartheta\right)\right]}$, quantity that has to be inserted in (13).

Really, intervals in (9) and (13) are, respectively, intervals for $\mathsf{E}\left[\widehat{p}_{jl}\left(\vartheta\right)\right]$ and $\mathsf{E}\left[\widehat{e}_j\left(\vartheta\right)\right]$, rather than for $p_{jl}\left(\vartheta\right)$ and $e_j\left(\vartheta\right)$; thus, they share the bias present in $\widehat{p}_{jl}$ and $\widehat{e}_j$, respectively (for the OCC case, see Ramsay 1991, p. 619).

4.2. Relative credibility curve

For a generic subject $S_i \in \mathcal{S}$, we can compute the relative likelihood

$$L_i\left(\vartheta\right) = \frac{\prod_{j=1}^{k}\prod_{l=1}^{m_j}\left[\widehat{p}_{jl}\left(\vartheta\right)\right]^{y_{ijl}}}{\max_{\vartheta}\left\{\prod_{j=1}^{k}\prod_{l=1}^{m_j}\left[\widehat{p}_{jl}\left(\vartheta\right)\right]^{y_{ijl}}\right\}} \tag{14}$$

of the various values of ϑ given his pattern of responses on the test and given the kernel-estimated OCCs. The function in (14) is also known as relative credibility curve (RCC; see, e.g, Lindsey 1973). The ϑ-value, say $\widehat{\vartheta}^{ML}$, such that $L_i\left(\vartheta\right) = 1$, is called the maximum likelihood (ML) estimate of the ability for S_j (see also Kutylowski 1997). It is interesting to note that, for tests with multiple-choice items, $\widehat{\vartheta}^{ML}$ is based not only on how many items were answered correctly, but also on whether the items answered correctly were difficult or easy, whether the items answered incorrectly were difficult or easy, whether the correctly answered items were of high quality or not, and whether the options chosen for incorrectly answered items were typical of stronger or weaker examinees. Thus, $\widehat{\vartheta}^{ML}$ makes use of much more information than the conventional total number of correct answers T, and will tend to be a more accurate estimate of ability. When there is a substantial difference between $\widehat{\vartheta}^{ML}$ and T, it is possible that the pattern of option choices for incorrectly-answered items gave important additional information about ability.

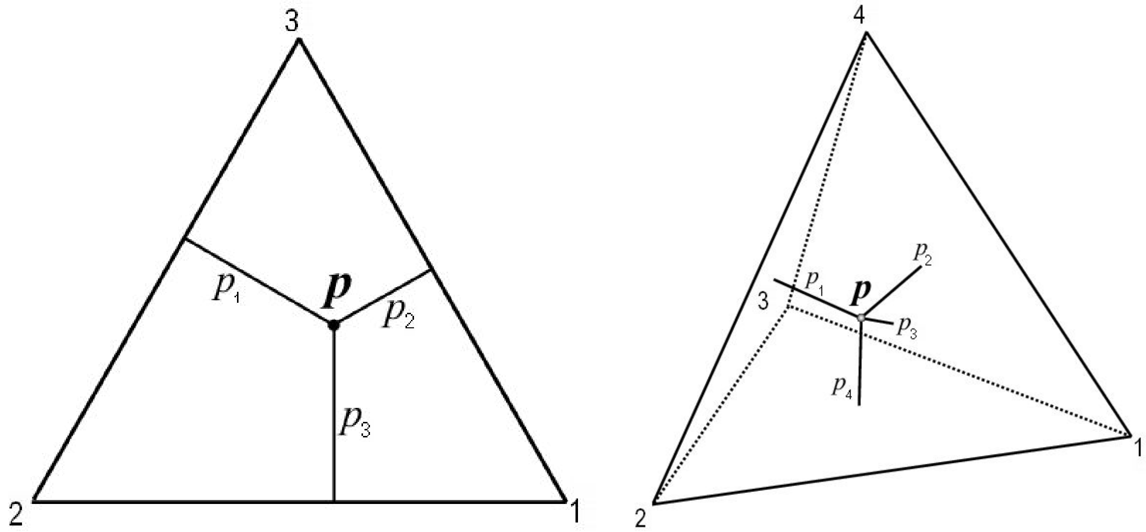
The relative likelihood $L_i\left(\vartheta\right)$ is generally a curve with only one maximum in $\widehat{\vartheta}^{ML}$, with concentration around $\widehat{\vartheta}^{ML}$ being an indication of its precision. Occasionally, the shape of (14) can have two maxima, and this indicates a response pattern giving a mixed message: the subject passed some difficult items, indicating high ability, and at the same time failed some easy items, suggesting lower ability. This can happen when the subject knows some part of the material well and another part poorly. The curve rightly reflects the resulting ambiguity about the subject's true ability.

Finally, as Kutylowski (1997) and Ramsay (2000) do, the obtained values of $\widehat{\vartheta}^{ML}$ may be used as a basis for a second step of a kernel smoothing estimation of the OCCs. This iterative process, consisting in cycling back the values of $\widehat{\vartheta}^{ML}$ into estimation, can clearly be repeated any number of times with the hope that each step refines or improves the estimates of ϑ. However, as the same Ramsay (2000) declares, for the vast majority of applications, no iterative refinement is really necessary, and the use of $\widehat{\vartheta}_i$ or $\widehat{\vartheta}_i^{ML}$ for ranking examinees works fine. This is the reason why we have not consider the iterative process in the package.

4.3. Probability simplex

With reference to a generic item $I_j \in \mathcal{I}$, the vector of probabilities $\widehat{\boldsymbol{p}}_j(\vartheta)$ can be seen as a point in the probability simplex $\mathbb{S}^{m_j}$, defined as the (m_j-1)-dimensional subset of the m_j-dimensional space containing vectors with nonnegative coordinates summing to one. As ϑ varies, since the assumptions of both smoothness and unidimenionality in the latent trait, $\widehat{\boldsymbol{p}}_j(\vartheta)$ moves along a curve; the item analysis problem is to locate the curve properly within the simplex. On the other hand, the estimation problem for S_i is the location of its position along this curve.

A convenient way of displaying points in $\mathbb{S}^3$ is represented by the *reference triangle* in Figure 1(a), an equilateral triangle, with vertices 1, 2, 3, having unit altitude (see Aitchison 2003, pp. 5–6). For any point $\boldsymbol{p}$ in the triangle 123 the perpendiculars p_1, p_2, p_3 from $\boldsymbol{p}$ to the sides



(a) Triangle (b) Tetrahedron

Figure 1: Convenient way of displaying a point in the probability simplex $\mathbb{S}^{m_j}$ when $m_j = 3$ (on the left) and $m_j = 4$ (on the right).

opposite to the vertices 1, 2, 3 satisfy

$$p_l \geq 0,\ l = 1, 2, 3,\ \text{and } p_1 + p_2 + p_3 = 1. \tag{15}$$

Since there is a unique point in triangle 123 with perpendicular values p_1, p_2, p_3, there is a one-to-one correspondence between $\mathbb{S}^3$ and points in triangle 123, and so we have a simple

means of representing the vector of probabilities $\widehat{\boldsymbol{p}}_j(\vartheta)$ when $m_j = 3$. In such a representation we may note that the three inequalities in (15) are strict if and only if the point lies in the interior of triangle 123. Also, the larger a component p_l is, the further the point is away from the side opposite the vertex l. Moreover, vectors (p_1, p_2, p_3) with two components, say p_2 and p_3, in constant ratio are represented by points on a straight line through the complementary vertex 1. For 4-dimensional vectors of probabilities we have to move into the 3-dimensional space to obtain a picture of $\mathbb{S}^4$ via a *regular tetrahedron* 1234 of unit altitude (see Aitchison 2003, pp. 8–9) taking the place of the reference triangle. In Figure 1(b) the probabilities p_l corresponds to the perpendicular from the point $\boldsymbol{p}$ to the triangular face opposite the vertex l. Note that for items with more than four options there is no satisfactory way of obtaining a visual representation of the corresponding probability simplex; nevertheless, we can perform a partial analysis which focus attention on some options for that item.

Finally note that, as discussed in Section 2, in practice only the values of the functions $\widehat{\boldsymbol{p}}_j$ are determined from the data while, by contrast, only the rank order of their arguments are known. Thus, one would like a display of the variation in the probability values $\widehat{\boldsymbol{p}}_j$ across subjects that tends to hide the role of the argument or domain variable ϑ. This is precisely the purpose of the probability simplex.

5. Package KernSmoothIRT in use

What follows is an illustration of the capabilities of the **KernSmoothIRT** package. The examples will highlight some of the more important functions, options and diagnostic plots. The examples are meant to be illustrative, not exhaustive.

5.1. Data input

The first tutorial will walk-through an analysis of a set of *multiple-choice items* while the second will walk-through a set of *rating scale items*. For either data type, the `ksIRT` function will perform the kernel smoothing. This function requires `responses` as well as a specification of the items type using the `scale` argument. Basic weighting of the items is governed by the `key` option while more complicated structures can be obtained via the `weights` argument. In particular, the `responses` argument must be a $(n \times k)$-matrix, with a row for each subject in $\mathcal{S}$ and a column for each item in $\mathcal{I}$, containing the selected option numbers. The `scale` argument indicates whether the items are multiple-choice, scale or a mixture of the two. The `key` argument must be a vector containing the correct response to each of the items in the case of multiple-choice, or the highest scale-level option in the case of rating scale items. When `key` is provided, a multiple choice response is scored correct or incorrect while a rating scale option is scored according to its corresponding number. For more complicated scoring schemes, such as partial credit, the user can input a list of weights for each item using the `weights` argument (see the help for details).

The user can also select the q evaluation points of Section 3.1, the ranking distribution F of Section 2, the type of kernel function K and the kernel bandwidth to input into the `ksIRT` function, though it will choose defaults if unspecified. In particular, by specifying `theta` or `nval` options, the user can respectively select the points, or their number q, at which to evaluate the OCCs. The default is data dependent, but can be overridden if the user would like more points or different limits for consistent comparisons across tests. Regarding F,

altering the `enumerate` option will allow for different distributions. The selection of a kernel function and kernel bandwidth are important choices as well. The `kernel` option allows for a Gaussian, uniform or quadratic kernel (Gaussian is chosen by default). The `bandwidth` option by default is specified according to the rule of thumb in equation (7). The user may input a numerical vector of bandwidths for each item to experiment with different levels of smoothing, or the user may input `bandwidth="CV"` to obtain cross-validation estimation of h_j, $j = 1, \ldots, k$, as described in Section 3.2.

Another consideration for the user is how to treat missing values. The option `miss`, of the function `ksIRT`, governs this aspect. The default, `miss="category"`, treats missing values as an option value themselves with zero weight. In this case, the OCC of the missing value will be added, and plotted, for the corresponding item. Also, it is possible to treat missing values as a category, but specify a non-zero weight with the `NAweight` option. Other choices impute the missing values according to some discrete probability distributions taking values on $\{1, \ldots, m_j\}$, $j = 1, \ldots, k$. In particular, by specifying `miss="random.unif"`, each missing value for the generic item $I_j \in \mathcal{I}$ is substituted with a value randomly generated from a discrete uniform distribution while, with `miss="random.multinom"`, each missing value for I_j is substituted with a number randomly generated from a multinomial distribution with probabilities equal to the frequencies amongst the non-missing responses to that item. Finally, the option `miss="omit"` will delete from the data set all the subjects with at least an omitted answer. The tools described in this section are not meant to be exhaustive or representative of all the capabilities of the **KernSmoothIRT** package. For further examples and descriptions of other analytical plots available, as well as other kernel smoothing options available, consult the `ksIRT` help page within the package.

5.2. Psych 101

The first tutorial uses the Psych 101 dataset included in the **KernSmoothIRT** package. This dataset contains the responses of $n = 379$ students, in an introductory psychology course, to $k = 100$ multiple choice items, each with $m_j = 4$ options as well as a key. These data were also analyzed in Ramsay and Abrahamowicz (1989) and in Ramsay (1991).

To begin the analysis, create a `ksIRT` object. This step performs the kernel smoothing and prepares the object for analysis using the many types of plots available.

```
R> data("Psych101")
R> Psych1 <- ksIRT(responses=Psychresponses, key=Psychkey, scale="nominal")
R> Psych1
```

```
    Item Correlation
1      1  0.23092838
2      2  0.09951663
3      3  0.19214764
.      .          .
.      .          .
.      .          .
99    99  0.01578162
100  100  0.24602614
```

The command `data("Psych101")` loads both `Psychresponses` and `Psychkey`. The function `ksIRT` produces kernel smoothing estimates using, by default, a Gaussian distribution F (`enumerate=list("norm",0,1)`), a Gaussian kernel function K (`kernel="gaussian"`), and the rule of thumb (7) for the bandwidths. The last command, `Psych1`, prints the point-polyserial correlations, traditional descriptive measures of items performance given by the correlation between each dichotomous/polythomous item and the total score (see Olsson, Drasgow, and Dorans 1982, for details).

Once the `ksIRT` object `Psych1` is created, plots become available to analyze each item, subject and the overall test. There are sixteen plots available to evaluate the test through the `plot` function by altering the `plottype` option.

OCCs

The code

```
R> plot(Psych1, plottype="OCC", item=c(24,25,92,96))
```

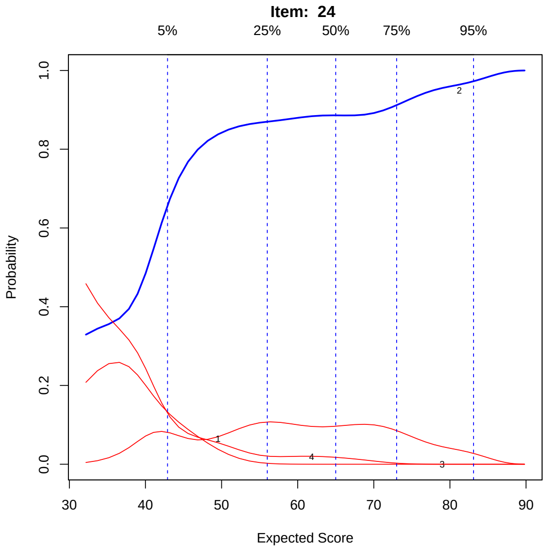
produces the OCCs for items 24, 25, 92, and 96 displayed in Figure 2. The correct options, for multiple-choice items like these, are displayed in green and the incorrect options in red. The specification `axistype="scores"` uses the expected total score (10) as display variable on the x-axis; the expected score is a transformation of the trait level to the number of items that a subject of that trait level would, on average, answer correctly. The vertical dashed lines indicate the scores (or quantiles if `axistype="distribution"`) below which 5%, 25%, 50%, 75% and 95% of subjects fall. Since the argument `miss` has not been specified, by default the "missing category" is plotted as an additional OCC (`miss="category"`), as we can see from Figure 2(b) and Figure 2(d) which refer to items with 2 and 1 nonresponses, on 379 cases.

The OCC plots in Figure 2 show four very different items. Globally, apart from item 96 in Figure 2(d), the other items appear to be monotone enough. Item 96 is problematic for the Psych 101 instructor as subjects with lower trait levels are more likely to select the correct option than higher trait level examinees. In fact, examinees with expected scores of 90 are the least likely to select the correct option. Perhaps the question is misworded or it is testing the wrong concept. On the contrary, items 24, 25, and 92, do a good job in differentiating between subjects with low and high trait levels. In particular item 24, in Figure 2(a), displays an high discriminating power for subjects with expected scores near 40, and a lower one for examinees with expected scores greater than 50 that have the same probability of selecting the correct option regardless of their expected score. Item 25 in Figure 2(b) is also a good item, only the top students are able to recongize option 3 as incorrect; option 3 was selected by about 30.9% of the test takers, or about 72.7% of those who answered incorrectly. Note also that, for subjects with expected scores below about 58, option 3 constitutes the most probable choice. Finally, item 92 in Figure 2(c), aside from being monotone, is also easy since a subject with expected score of about 30 already has a 70% chance of selecting the correct option; only a few examinees are consequently interested to the incorrect options 1, 3, and 4.

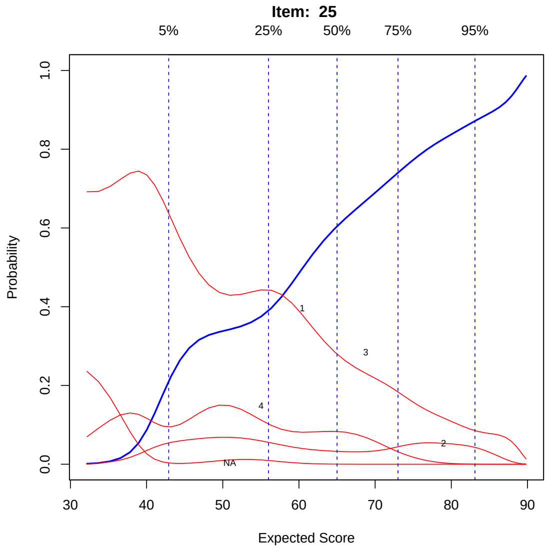
ICCs

Through the code

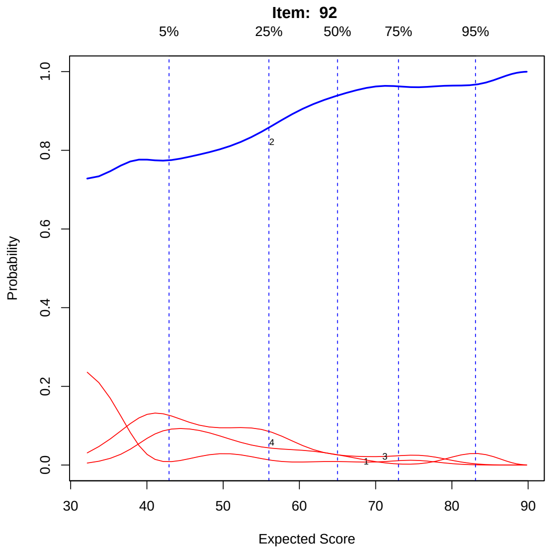
```
R> plot(Psych1, plottype="ICC", item=c(24,25,92,96))
```



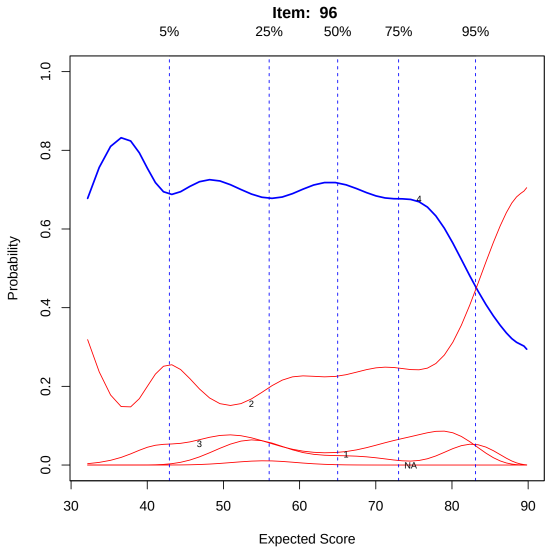
(a) Item 24



(b) Item 25



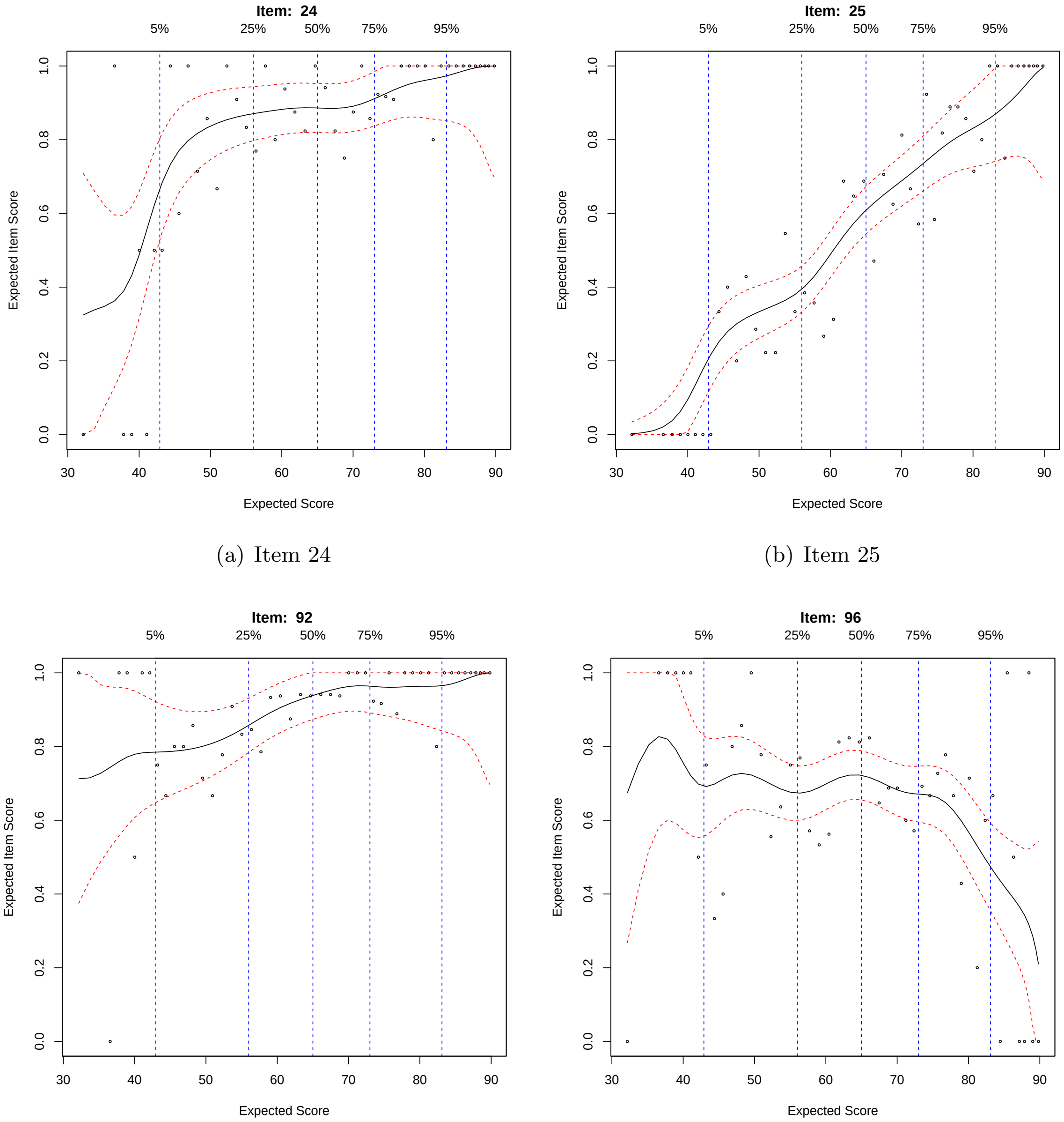
(c) Item 92



(d) Item 96

Figure 2: OCCs for items 24, 25, 92, and 96 of the introductory psychology exam.

we obtain, for the same set of items, the ICCs displayed in Figure 3. As said before, due to the 0/1 weighting scheme, in the case of multiple choice items, the ICC is the same as the OCC (shown in green in Figure 2) for the correct option. ICCs by default show the 95% pointwise confidence intervals (dashed red lines) illustred in Section 3.3. Via the argument `alpha`, confidence intervals can be removed entirely (`alpha=FALSE`) or changed by specifying a different value. In this example, relatively wide confidence intervals, for expected total scores at extremely high or low levels, are obtained. This is due to the fact that there are less data for estimating the curve in these regions and thus there is less precision in the estimates.



(a) Item 24

(b) Item 25

(c) Item 92

(d) Item 96

Figure 3: ICCs, and corresponding 95% pointwise confidence intervals (dashed red lines), for items 24, 25, 92, and 96 of the introductory psychology exam. Grouped subject scores are displayed as points.

Finally, the points on the ICC plots show the grouped subject scores illustrated in Section 3.1.

Probability simplex plots

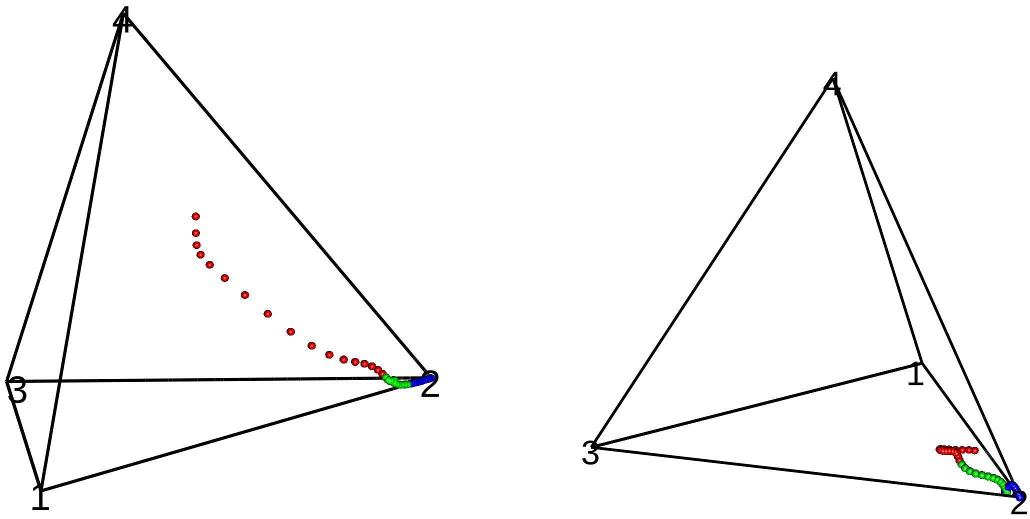
To complement the OCCs, the package includes triangle and tetrahedron (simplex) plots that, as illustrated in Section 4.3, synthesize the OCCs. When these plots are used on

items with more than 3 or 4 options (including the missing value category), only the options corresponding to the 3 or 4 highest probabilities will be shown; naturally, these probabilities are normalized in order to allow the simplex representation. This seldom loses any real information since experience tends to show that in a very wide range of situations people tend to eliminate all but a few options.

The tetrahedron is the natural choice for the items 24 and 92, characterized by four options and without "observed" missing responses; for these items the code

```
R> plot(Psych1, plottype="tetrahedron", items=c(24,92))
```

generates the tetrahedron plots displayed in Figure 2. These plots may be manipulated with



(a) Item 24 (b) Item 92

Figure 4: Probability tetrahedrons for two items of the introductory psychology exam. Low trait levels are plotted in red, medium in black and high in blue.

the mouse or keyboard as any other plot created with the package **rgl**. Inside the tetrahedron there is a curve constructed from a number of points. As said before, each point corresponds to a trait level. In particular, low, medium and high trait levels are identified by red, green and blue points, respectively. Considering this ordering in the trait level, it is possible to make some considerations.

- A basic requirement of a reasonable test item is that the sequence of points terminates at or near the correct answer. In these terms, as can be noted in Figure 4(a) and Figure 4(b), items 24 and 92 satisfy this requirement since the sequence of points moves toward the correct option, which is O_2 for both the items.

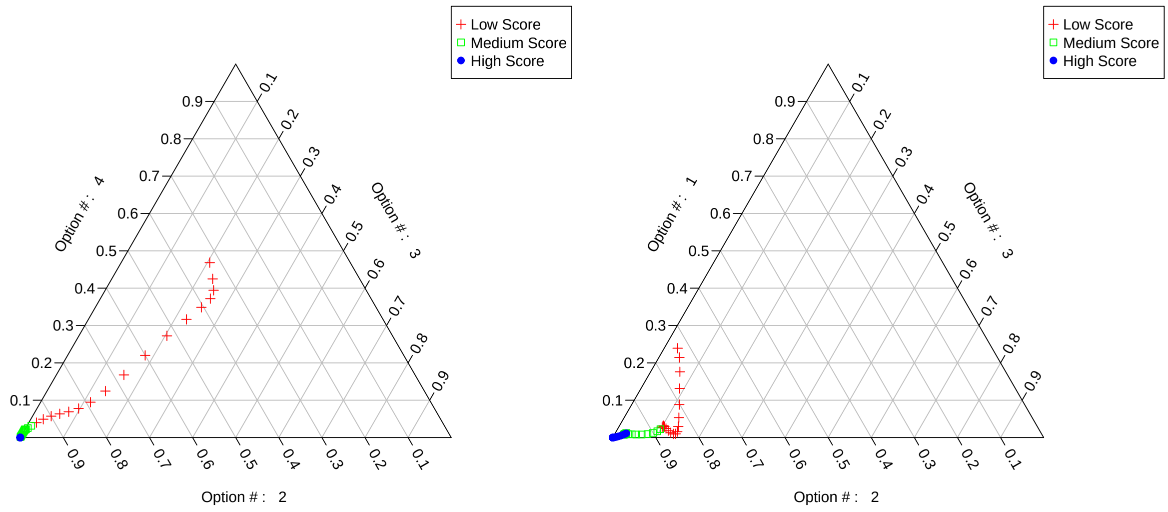
- The length of the curve is very important. The individuals with the lowest trait levels should be far from those with the highest. Item 24, in Figure 4(a), is a fairly good example. By contrast, very easy test items, such as item 92 in Figure 4(b), have very short curves concentrated close to the correct answer, with only the worse students showing a slight tendency to choose a wrong answer.

- The relative spacing of the points indicates the speed at which probabilities of choice changes. In these terms, see the contrast between items 24 and 92, in Figure 2, among the worst students.

Naturally, all these considerations are also obvious from Figure 2(a) and Figure 3(a). For the same items, the code

```
R> plot(Psych1, plottype="triangle", items=c(24,92))
```

produces the triangle plots displayed in Figure 5. For example, from Figure 5(a) we can see



(a) Item 24 (b) Item 92

Figure 5: Probability triangles for two items of the introductory psychology exam.

that the set of three most chosen options (O_2, O_3 and O_4), O_2 have much higher probability of selection while the other two are characterized by almost the same probability of selection since the sequence of points approximately lies on the bisector of the angle associated to O_2.

Principle component analysis

By performing a principal component analysis (PCA) of the ICCs at each point of evaluation, the **KernsmoothIRT** package provides a way for simultaneously compare items and show the relationships among them. In particular, the code

```
R> plot(Psych1, plottype="PCA")
```

produces the graphical representation in Figure 6. In the interior plot we have the graphical representation of the first two components obtained by a PCA on the values of the ICCs at each evaluation point $\vartheta_1, \ldots, \vartheta_s, \ldots, \vartheta_q$. In detail, the average ICC is preliminary calculated across items and subtracted from each ICC; in other words, the PCA is carried out on the centered ICCs. The dashed lines on the interior plot show the average item for each component. A first glance to this plot shows that:

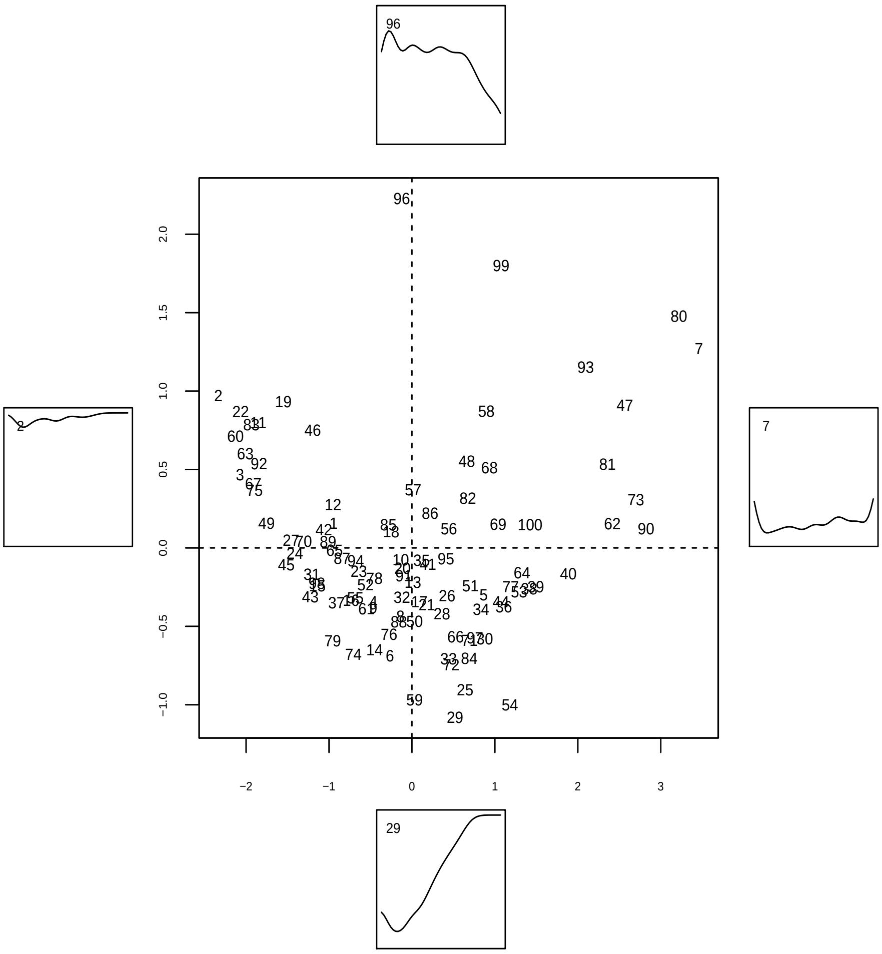


Figure 6: First two principal components for the introductory psychology exam. In the interior plot, numbers are the identifiers of the items. The vertical component represents discrimination, while the horizontal one difficulty. The small plots show the ICCs for the most extreme items for each component.

- the first principal component, plotted as the horizontal axis, represents item difficulty, since the most difficult items are placed on the right and the easiest ones on the left. The small plots on the left and on the right show the ICCs for the two extreme items with respect to this component and help the user in identifying the axis-direction with respect to difficulty (from low to high or from high to low). Here, item 7 shows high difficulty, as test takers of all ability levels receive a low score while item 2 is extremely easy.

- the second principal component, on the vertical axis, corresponds to item discriminability since low items tend to have an high positive slope while low items tend to be have an high negative slope. Also in this case, the small plots on the bottom and on the top show the ICCs for the two extreme items with respect to this component and help the user in identifying the axis-direction with respect to discrimination (from low to high

or from high to low). Here, item 29 discriminates very well whereas item 96 does not, it negatively discriminates.

We also note that, items 96 and 99 are outliers, since they possess a very negative discriminability, while items 7 and 80 are outliers because they are very difficult. Concluding, the principal components plot tends to be a useful overall summary of the composition of the test. Figure 6 is fairly typical of most academic tests and it is also usual to have only two dominant principal components reflecting item difficulty and discrimination.

RCCs

The RCCs shown in Figure 7 are obtained by the command

```
R> plot(Psych1, plottype="credibility", subjects=c(33,92,111,183))
```

In each plot, the red line shows the subject's actual score T.

For both the subjects considered in Figure 7(a) and Figure 7(b), there is a substantial agreement between the maximum of the RCC, $\widehat{\vartheta}^{ML}$, and T. Nevertheless, there is a difference in terms of the precision of the ML-estimates; for subject 183 the RCC is indeed more spiky, denoting an higher precision. In Figure 7(c) there is a substantial difference between $\widehat{\vartheta}^{ML}$ and T. This indicates that the correct and incorrect answers of this subject are more consistent with a lower score than they are with the actual score received. Finally, in Figure 7(d), although there is a substantial agreement between $\widehat{\vartheta}^{ML}$ and T, a small but prominent bump is present in the right part of the plot. Although subject 33 is well represented by his total score, he passed some, albeit few, difficult items and this may lead to think that he is more able than T_{33} suggests.

The commands

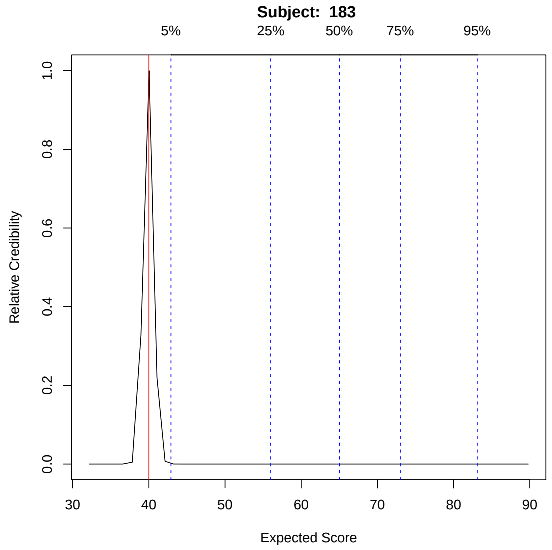
```
R> Psych1$scoresbysubject

[1] 74 56 89 70 56 57 ...

R> Psych1$subMLE

[1] 73.48316 59.06626 89.00686 67.47167 57.71787 55.03844 ...
```

allow us to evaluate the differences between the values of T_i and $\widehat{\vartheta}_i^{ML}$, $i = 1, \ldots, n$.

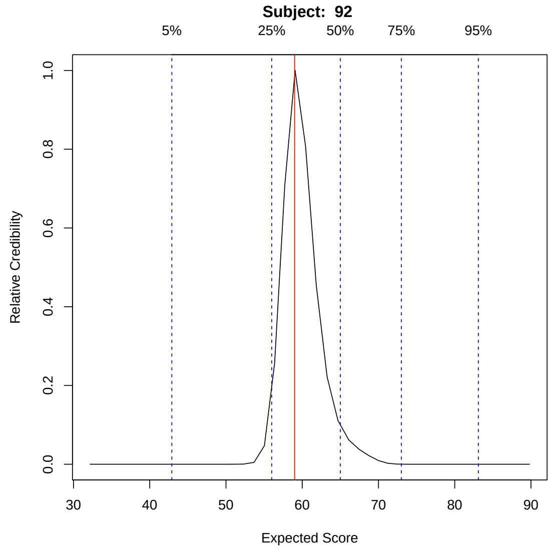
Test summary plots

The **KernSmoothIRT** package also contains many analytical tools to assess the test overall. Figure 8 shows a few of these, obtained via the code

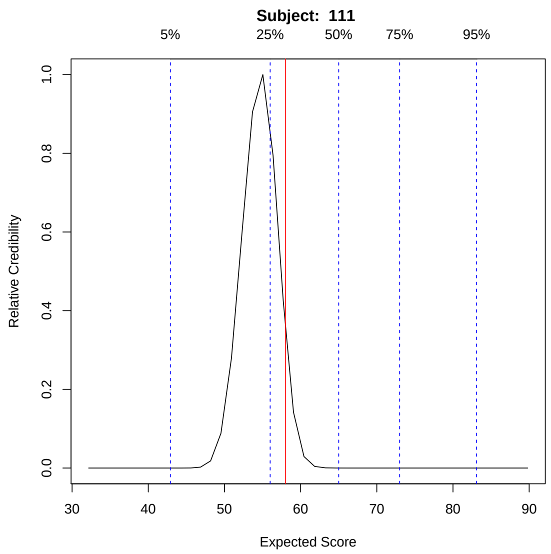
```
R> plot(Psych1, plottype="expected")
R> plot(Psych1, plottype="sd")
R> plot(Psych1, plottype="density")
```



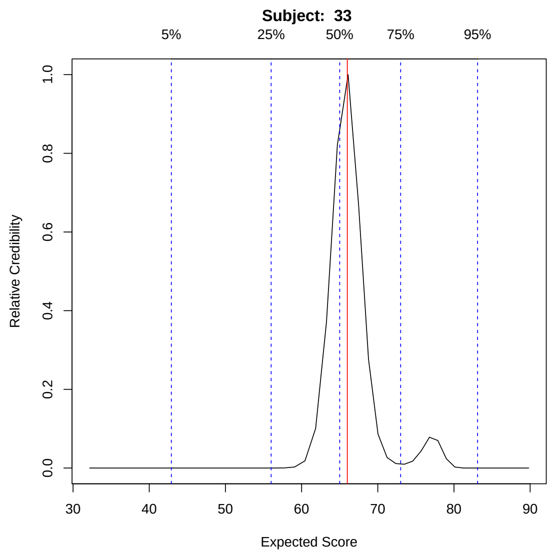
(a) Subject 183



(b) Subject 92



(c) Subject 111

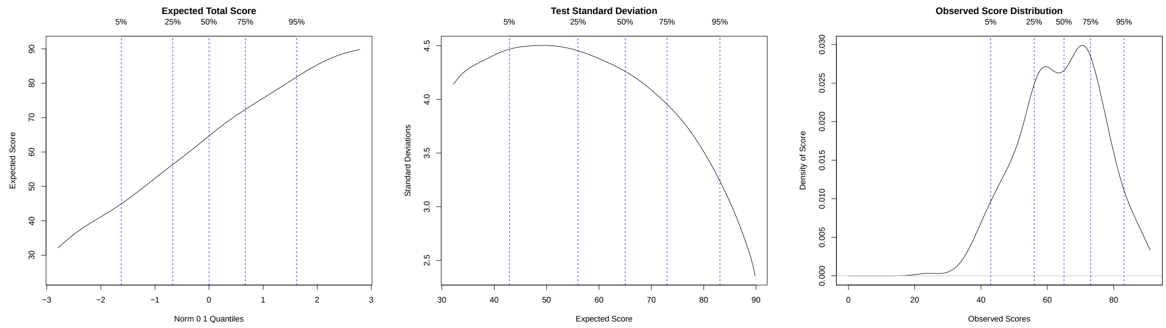


(d) Subject 33

Figure 7: RCCs for some subjects. The vertical red line shows the actual score the subject received.

Figure 8(a) shows the so-called Test Characteristic Function (TCF), which is the transformation of the quantiles of F into the expected scores. The TCF, for the Psych 101 dataset, is nearly linear. Note that, in the nonparametric context, the TCF may be non-monotonic due to either ill-posed items or random variations. In the latter case, a slight increase of the bandwidth may be advisable.

The total score T, for subjects having a particular value ϑ, is a random variable, in part because different examinees, or even the same examinee on different occasions, cannot be



(a) Test Characteristic Function (b) Standard Deviation (c) Density

Figure 8: Test summary plots for the introductory psychology exam.

expected to make exactly the same choices. The standard deviation of these values, graphically represented in Figure 8(b), is therefore also a function of ϑ, denoted by $\sigma_T(\vartheta)$. Figure 8(b) indicates that $\sigma_T(\vartheta)$ reaches the maximum for examinees at around a total score of 50, where it is about 4.5 items out of 100. This translates into 95% confidence limits of about 41 and 59 for a subject getting 50 items correct. Low proficiency subjects have a relatively high standard deviation in their scores relative to high proficiency subjects.

Figure 8(c) shows a kernel density estimate of the distribution of T. Although such distribution is commonly assumed to be "bell-shaped", from this plot we can note as this assumption is strong for these data. In particular, a negative skewness can be noted which is a consequence of the exam having relatively more easy items than hard ones. Moreover, bimodality is evident with modes at $T = 60$ and $T = 70$.

5.3. Voluntary HIV-1 counseling and testing efficacy study group

It is often useful to explore if, for a specific item on a questionnaire or test, its characteristic curves differ when estimated on two or more different groups of subjects, commonly formed by gender or ethnicity. This is called Differential Item Functioning (DIF) analysis in the psychometric literature. In particular, DIF occurs when subjects with the same ability but belonging to different groups have a different probability of choosing a certain option. DIF can properly be called *item bias* because the characteristic curves of an item should depend only on ϑ, and not directly on other person factors. Zumbo (2007) offers a recent review of various DIF detection methods and strategies.

The **KernSmoothIRT** package allows for a nonparametric graphical analysis of DIF, based on kernel smoothing methods. To illustrate this analysis, we will use data coming from the Voluntary HIV Counseling and Testing Efficacy Study, conducted in 1995-1997 by the Center for AIDS Prevention Studies (see The Voluntary HIV-1 Counseling and Testing Efficacy Study Group 2000a,b, for details). This study was concerned with the effectiveness of HIV counseling and testing in reducing risk behavior for the sexual transmission of HIV. To perform this study, $n = 4292$ persons were enrolled. The whole dataset – downloadable from `http://caps.ucsf.edu/research/datasets/`, which also contains other useful survey details – reported 1571 variables for each participant. As part of this study, respondents were surveyed about their attitude toward condom use via a bank of $k = 15$ items. Respondents

were asked how much they agreed with each of the statements on a 4-point response scale, with 1="strongly disagree", 2="disagree more than I agree", 3="agree more than I disagree", 4="strongly agree"). Since 10 individuals omitted all the 15 questions, they have been preliminary removed from the used data. Moreover, given the ("negative") wording of the items I_2, I_3, I_5, I_7, I_8, I_{11}, and I_{14}, a respondent who strongly agreed with such statements was indicating a less favorable attitude toward condom use. In order to uniform the data, the score for these seven items was preliminary reversed. The dataset so modified can be direct loaded from the **KernSmoothIRT** package by the code

```
R> data("HIV")
R> HIV
```

```
     SITE GENDER AGE I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 I13 I14 I15
   1  Ken      F  17  4  1  1  4  1  2  4  4  4   4   3   4   1   2   4
   2  Ken      F  17  4  2  4  4  2  3  1  4  3   3   2   3   4   1   4
   3  Ken      F  18  4  4  4  4  4  1  4  4  4   1  NA   4   1  NA   4
   4  Ken      F  18  4 NA  1  4  1  1 NA  4  2   1   4   2   1   3   3
   5  Ken      F  18  4  1  1  3  1  1 NA  1  2   1   3   2   1   1   3
   6  Ken      F  18  4  4  4  4  4  1  3  1  2   1   4   2   3  NA   4
   .    .      .   .  .  .  .  .  .  .  .  .  .   .   .   .   .   .   .
   .    .      .   .  .  .  .  .  .  .  .  .  .   .   .   .   .   .   .
   .    .      .   .  .  .  .  .  .  .  .  .  .   .   .   .   .   .   .
4277  Tri      M  72  4  4  1  4  1  4  2  4  4   4   2   4   1   2   1
4278  Tri      M  72  4  4  4  4  4  4  4  4  4   1   1   4   1   4   4
4279  Tri      M  73  2  4  2  3 NA NA NA NA NA  NA  NA  NA  NA   1  NA
4280  Tri      M  76  4  4  1  4  1  1  1  4  4   4   4   4   1   1  NA
4281  Tri      M  79  4  4  1  4  1 NA  4 NA  4  NA  NA  NA  NA   1   4
4282  Tri      M  80  4 NA  4  4  1  4 NA NA NA   1  NA   4   1   4  NA
```

```
R> attach(HIV)
```

As it can be easily seen, the above data frame also contains the following person factors:

SITE = "site of the study" (Ken=Kenya, Tan=Tanzania, Tri=Trinidad)
GENDER = "subject's gender" (M=male, F=female)
AGE = "subject's age" (age at last birthday)

Each of these factors can be potentially used for a DIF analysis. These data have been also analyzed, through some well-known parametric models, by Bertoli-Barsotti, Muschitiello, and Punzo (2010) which also perform a DIF analysis. Part of this sub-questionnaire has been also considered by De Ayala (2003, 2009) with a Rasch Analysis.
The code below

```
R> HIVres <- ksIRT(HIV[,-(1:3)], key=HIVkey, scale="ordinal", miss="omit")
R> HIVres
```

```
   Item Correlation
1     1   0.2112497
```

```
2     2   0.4190828
3     3   0.4195175
4     4   0.2869221
5     5   0.4070306
6     6   0.3247448
7     7   0.4265602
8     8   0.4463606
9     9   0.4470928
10   10   0.2784581
11   11   0.4146700
12   12   0.3998968
13   13   0.3201667
14   14   0.2776529
15   15   0.3554800
```

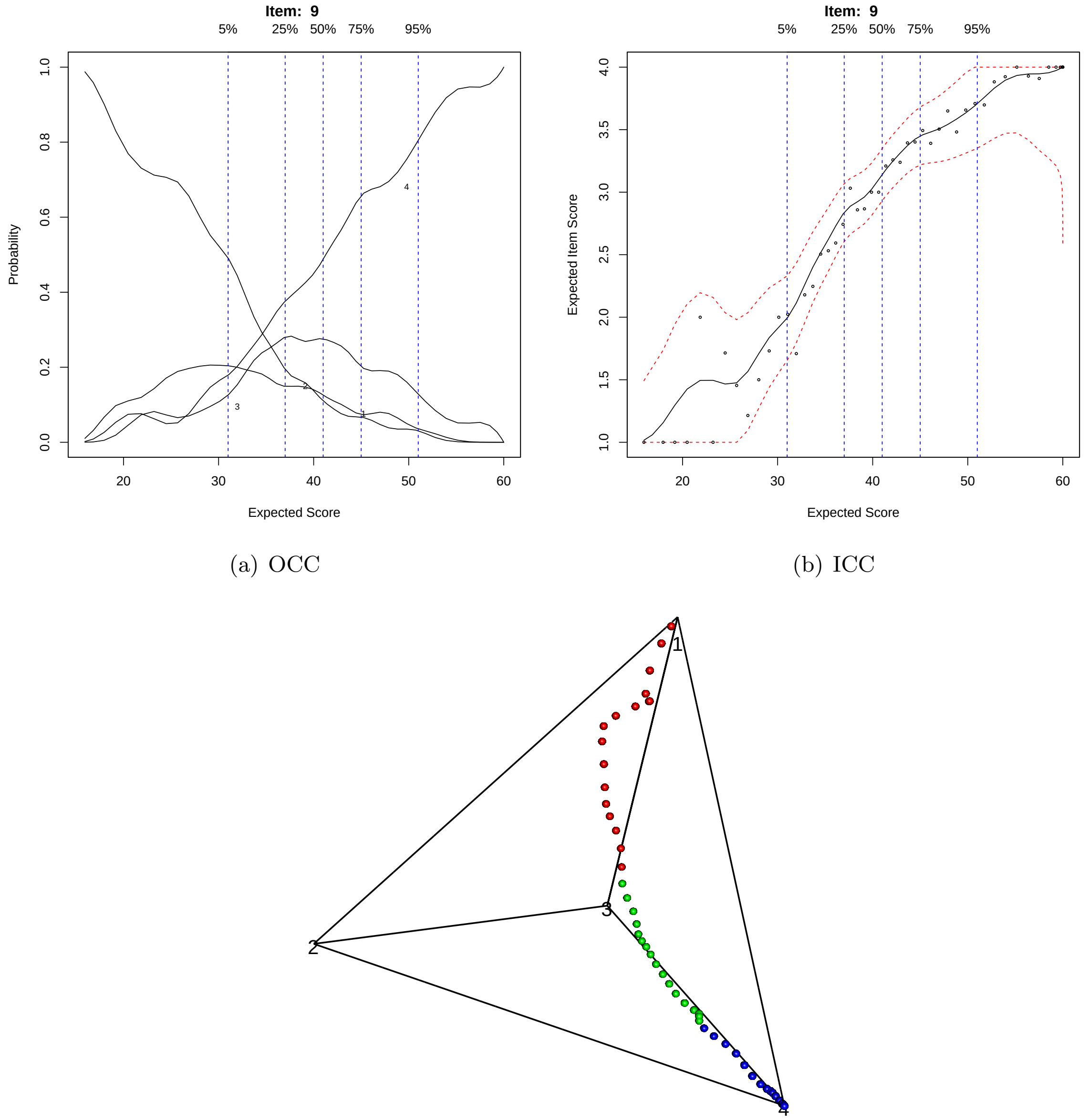
```
R> plot(HIVres, plottype="OCC", item=9)
R> plot(HIVres, plottype="ICC", item=9)
R> plot(HIVres, plottype="tetrahedron", item=9)
```

produces the plots displayed in Figure 9 for I_9. The option `miss="omit"` excludes from the nonparametric analysis all the subjects with at least an omitted answer, leading to a sample of 3473 respondents; the option `scale="ordinal"` specifies the rating scale nature of the items. Figure 9(a) displays the OCCs for the considered item. As expected, subjects with the smallest scores are choosing the first option while those with the highest ones are selecting the fourth option. Generally, as the total scores increase, respondents are approximately estimated to be more likely to choose an higher option and this reflects the typical behavior of a rating scale item. From the Figure 9(b), which shows the ICC for item 9, it may be observed how the expected item score climbs consistently as the total test score increases. Moreover, the ICC displays a fairly monotonic behavior that covers the entire range $[1, 4]$. Finally, Figure 9(c) shows the tetrahedron for item 9. It corroborates the good behavior of I_9 already seen in Figure 9(a) and Figure 9(b). The sequence of points herein, as expected, starts from (the vertex) O_1 and gradually terminates at O_4, passing from O_2 and O_3.

In the following, we provide an example of DIF analysis using the person factor `GENDER`. To perform the DIF analysis, a new `ksIRT` object must be created with the addition of the `groups` argument by which the different subgroups may be specified. In particular, the code

```
R> gr1 <- as.character(HIV$GENDER)
R> DIF1 <- ksIRT(res=HIV[,-(1:3)], key=HIVkey, scale="ordinal",
 +               groups=gr1, miss="omit")
R> plot(DIF1, plottype="expectedDIF", lwd=2)
R> plot(DIF1, plottype="densityDIF", lwd=2)
```

produces the plots in Figure 10. Figure 10(a) displays the QQ-plot between the distributions of the expected scores for males and females; if the performances of the two groups are about the same, the relationship will appear as a nearly diagonal line (a dotted diagonal line is plotted as a reference). Figure 10(b) shows the density functions for the two groups. Both



(a) OCC

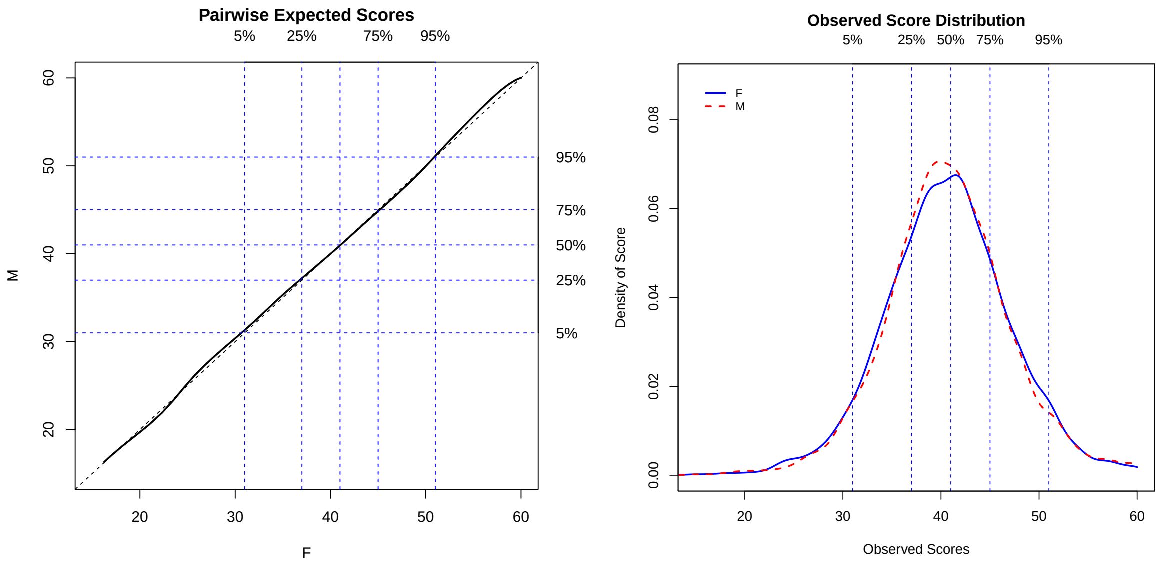
(b) ICC

(c) Tetrahedron

Figure 9: Item 9 from the voluntary HIV-1 counseling and testing efficacy study group.

plots confirm that there is a strong agreement in behavior of the two groups with respect to the test.

After this preliminary phase, the DIF analysis proceeds by considering the item by item group comparisons. Figure 11 displays the OCCs for the (rating scale) item I_3. These plots allow the user to compare the two groups at the item level. Lack of DIF is here manifested by OCCs being nearly coincident for all the four options. DIF may also be evaluated in terms of the expected scores of the groups, as displayed in Figure 12. This plot is obtained with the code



(a) Expected scores (b) Density

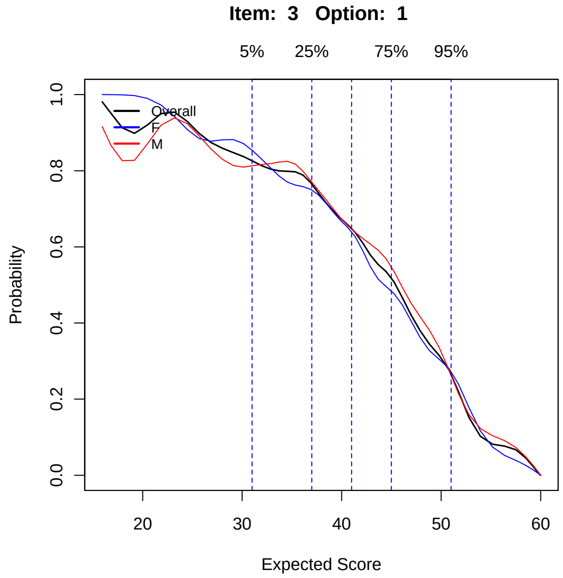
Figure 10: Behavior of male (`M`) and female (`F`) on the test. In the QQ-plot on the left, the dashed diagonal line indicates the reference situation of no difference in performance for the two groups; the horizontal and vertical dashed blue lines indicate the 5%, 25%, 50%, 75% and 95% quantiles for the two groups.

```
R> plot(DIF1, plottype="ICCDIF", cex=0.5, item=3)
```

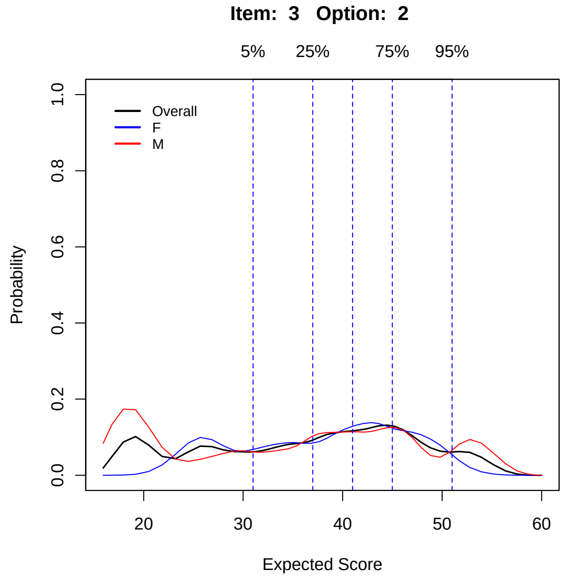
The different color points on the plot represent how individuals from the groups actually scored on the item. Although we have focused the attention only on I_3, similar results are obtained for all of the other items in $\mathcal{I}$, and this confirms as `GENDER` is not a variable producing DIF in this study. This result is confirmed in Bertoli-Barsotti *et al.* (2010). Note that, for both OCCs and ICCs, it is possible to add confidence intervals, through the `alpha` option. The code

```
R> gr2 <- as.character(HIV$SITE)
R> DIF2 <- ksIRT(res=HIV[,-(1:3)], key=HIVkey, scale="ordinal",
 +                  groups=gr2, miss="omit")
R> plot(DIF2, plottype="expectedDIF", lwd=2)
R> plot(DIF2, plottype="densityDIF", lwd=2)
```

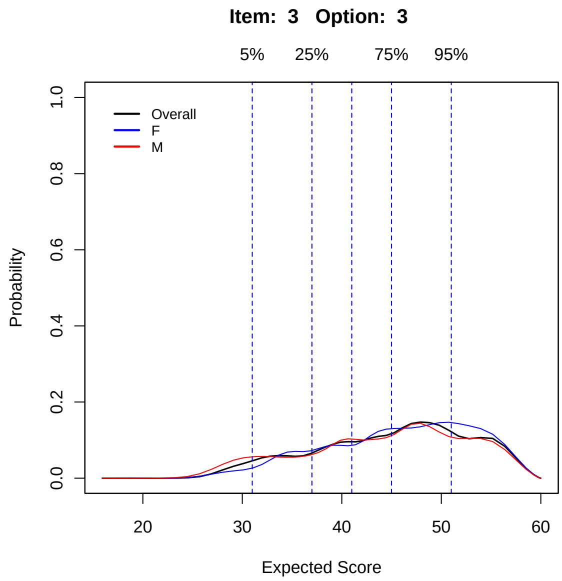
produces the plots in Figure 13, differentiate amongst subjects with different `SITE` levels. Among the 3473 subjects answering to all the 15 items, 987 come from Trinidad, 1143 come Kenya and 1346 from Tanzania. As highlighted by Bertoli-Barsotti *et al.* (2010), there are differences among these groups, and Figure 13 shows this. The three pairwise QQ-plots of the expected score distributions show that there is a slight dominance of people from Kenia over people from Trinidad (in the sense that people from Kenia have, in distribution, a slightly greater attitude toward condom use than people from Trinidad), and a large discrepancy between the performances of people from Tanzania relative to both other groups, as shown in



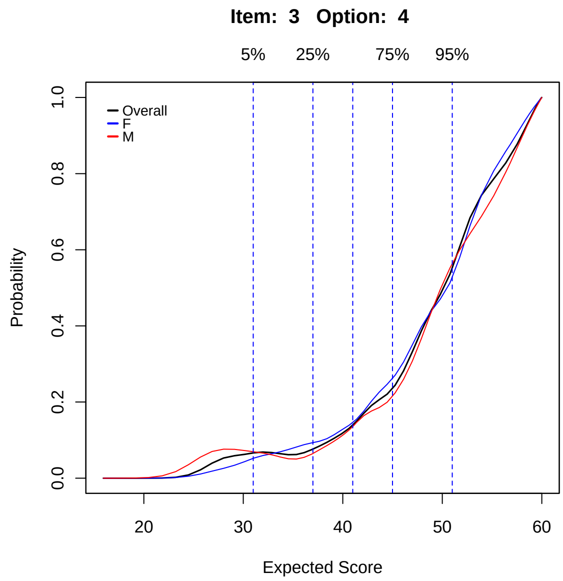
(a) Option 1



(b) Option 2



(c) Option 3



(d) Option 4

Figure 11: OCCs, for males and females, related to the item 3 of the voluntary HIV-1 counseling and testing efficacy study group. The overall OCCs are also superimposed.

Figure 13(a) Figure 13(b). The above dominance, and the peculiar behavior of people from Tanzania compared with the other countries, can be also noted by looking at the expected total score densities in Figure 13(d). Here, there is more higher variability in the total score for people from Tanzania. But what about DIF? The command

```
R> plot(DIF2, plottype="ICCDIF", item=c(6,11))
```

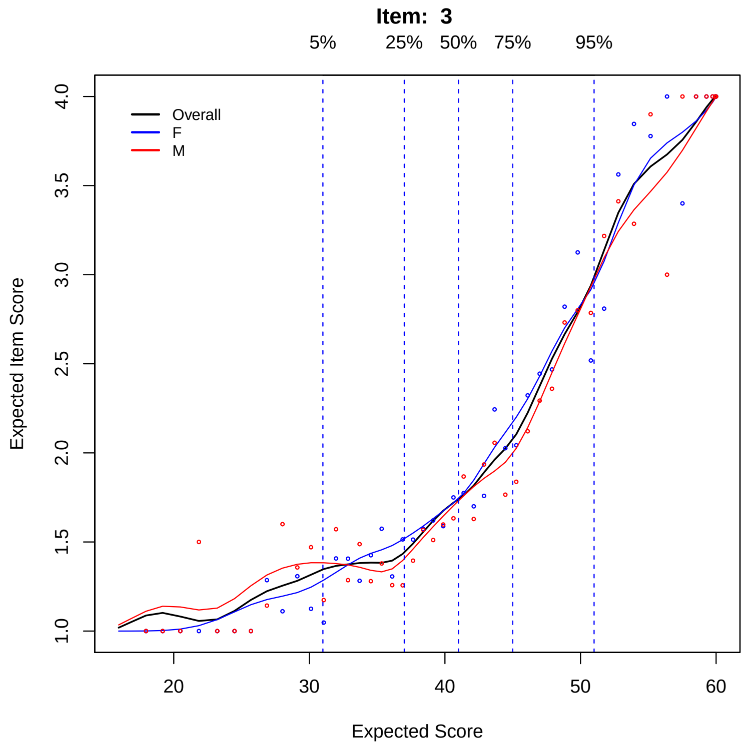


Figure 12: ICCs for male and female for item 3.

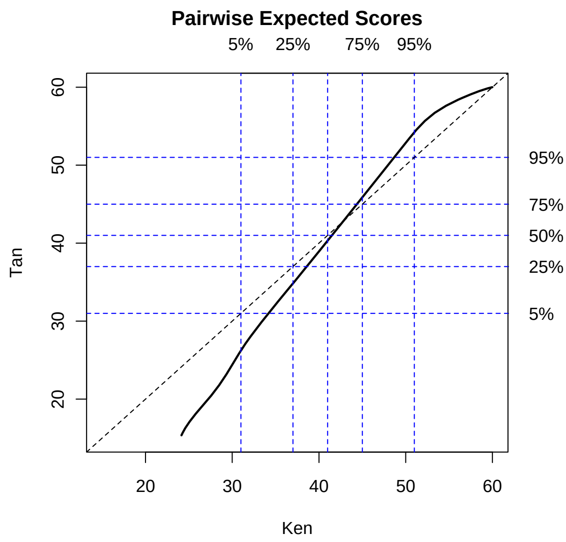
produces the ICCs in Figure 14, for I_6 and I_{11}. In both the plots we have a graphical indication of the presence of DIF for I_6 and I_{11}, and this confirms the results by Bertoli-Barsotti *et al.* (2010) that detect site-based DIF for these and other items in the test.

6. Conclusions

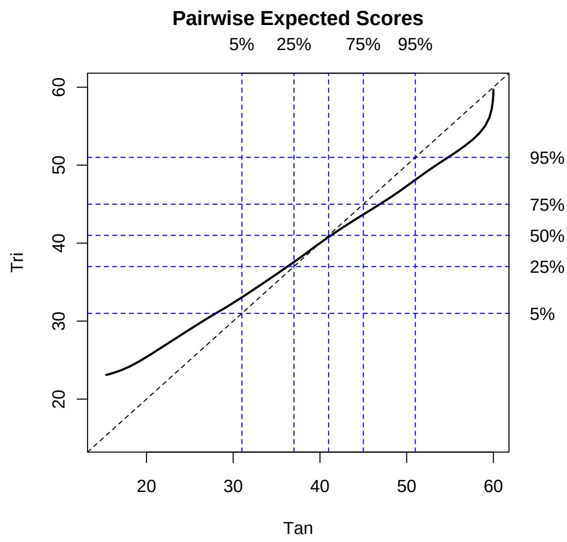
In this paper some theoretical as well as practical considerations over the use of kernel smoothing in IRT have been presented, with respect to the application of the **KernSmoothIRT** package for the R environment.

The advantages of nonparametric IRT modeling are well known. Ramsay (2000) recommends its application at least as an explorative tool, to guide the user over the choice of an appropriate parametric model. While most current IRT analyses are conducted with parametric models; quite often the assumption underling parametric IRT modeling are not preliminarily checked. One reason for this may be the lack, apart from TestGraf, of available software. TestGraf has set a milestone on this field, being it the first computer program to implement a kernel smoothing approach at IRT and the prominent software used over the years. With respects to TestGraf, our package has the major advantage of running within the R environment. Users do not have to export their results into another piece of software, in order to perform non-standard data analysis or to produce personalized plots. Also, within the same environment, the user may perform parametric IRT using one of several packages available.

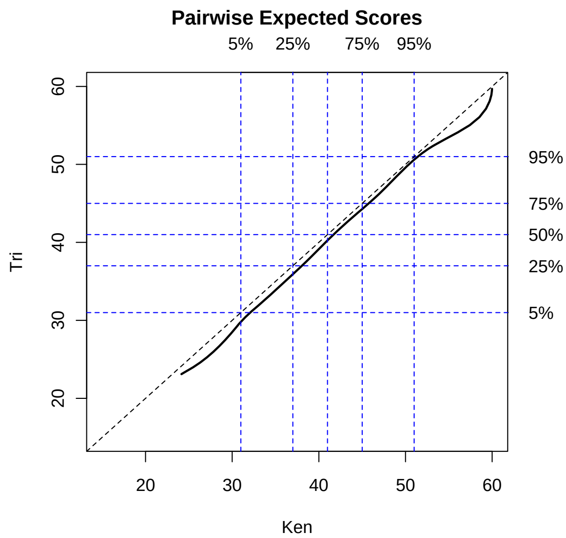
We believe that **KernSmoothIRT** may prove useful to lecturers and psychologists developing



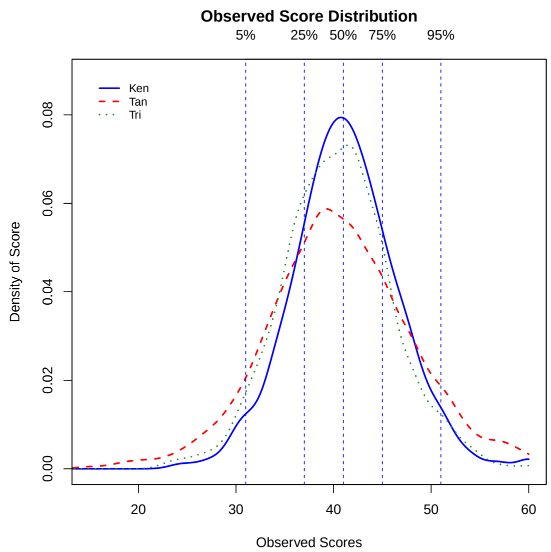
(a) QQ-plot (`Ken` vs. `Tan`)



(b) QQ-plot (`Tan` vs. `Tri`)



(c) QQ-plot (`Ken` vs. `Tri`)

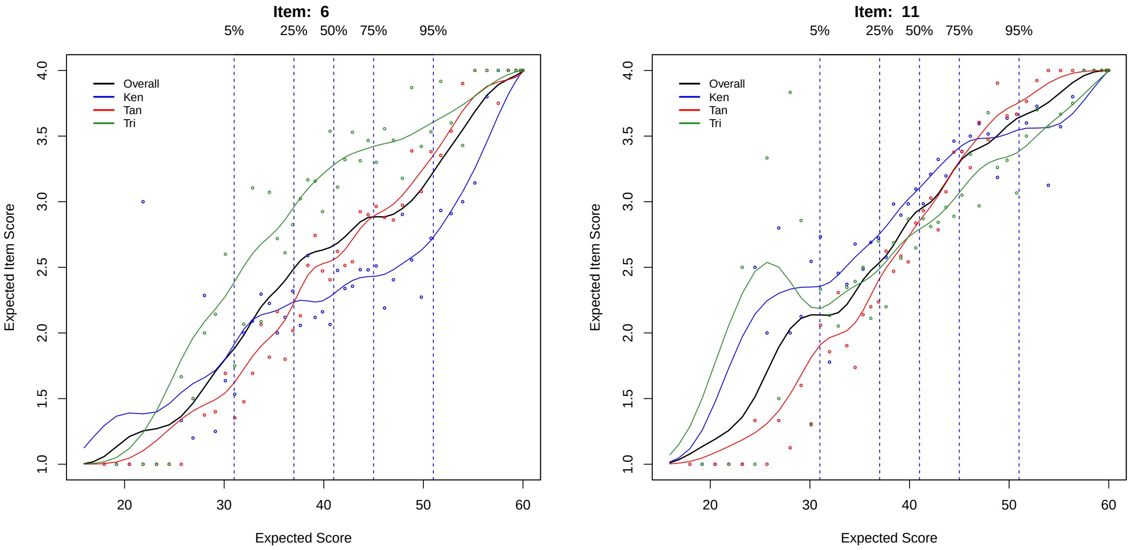


(d) Total score density

Figure 13: Behavior of people from Kenya (`Ken`), Tanzania (`Tan`), and Trinidad (`Tri`), on the test. In all the pairwise QQ-plots, the dashed diagonal line indicates the reference situation of no difference in performance for the two groups; the horizontal and vertical dashed blue lines indicate the 5%, 25%, 50%, 75% and 95% quantiles for the two groups.

questionnaires for test diagnostics such as spotting ill-posed questions and formulating more plausible wrong options. Also, in the paper we show how **KernSmoothIRT** can be used to do DIF analysis, by graphically displaying differences among groups of subjects in terms of how they respond to items.

Future work will consider extending the package by allowing for kernel smoothing estimation



(a) Item 6 (b) Item 11

Figure 14: ICCs, for people from Kenya (`Ken`), Tanzania (`Tan`), and Trinidad (`Tri`), for items 6 and 11 of the voluntary HIV-1 counseling and testing efficacy study group.

of test and item information functions too. Although well-established in parametric IRT, information functions present serious statistical problems – as also underlined by Ramsay (2000, p. 66) – in our context. Currently available nonparametric-based IRT programs such as TestGraf estimate test and item information functions based on parametric OCCs.

References

Aitchison J (2003). *The Statistical Analysis of Compositional Data*. Blackburn Press, Caldwell.

Altman NS (1992). "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression." *The American Statistician*, **46**(3), 175–185.

Baker FB, Kim S (2004). *Item Response Theory: Parameter Estimation Techniques*. Marcel Dekker, New York.

Bartholomew D (1983). "Latent Variable Models for Ordered Categorical Data." *Journal of Econometrics*, **22**(1–2), 229–243.

Bartholomew D (1988). "The Sensitivity of Latent Trait Analysis to Choice of Prior Distribution." *British Journal of Mathematical and Statistical Psychology*, **41**(1), 101–107.

Bertoli-Barsotti L, Muschitiello C, Punzo A (2010). "Item Analysis of a Selected Bank from the Voluntary HIV-1 Counseling and Testing Efficacy Study Group." *Technical Report 1*, Dipartimento di Matematica, Statistica, Informatica e Applicazioni (Lorenzo Mascheroni), Università degli studi di Bergamo.

Bock RD, Aitkin M (1981). "Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm." *Psychometrika*, **46**(4), 443–459.

Bock RD, Lieberman M (1970). "Fitting a Response Model for n Dichotomously Scored Items." *Psychometrika*, **35**(2), 179–197.

Chang HH, Mazzeo J (1994). "The Unique Correspondence of the Item Response Function and Item Category Response Functions in Polytomously Scored Item Response Models." *Psychometrika*, **59**(3), 391–404.

De Ayala R (2003). "The Effect of Missing Data on Estimating a Respondent's Location Using Ratings Data." *Journal of Applied Measurement*, **4**(1), 1–9.

De Ayala R (2009). *The Theory and Practice of Item Response Theory*. Guilford Press, New York.

de Leeuw J, Mair P (2007). "An Introduction to the Special Volume on "Psychometrics in R"." *Journal of Statistical Software*, **20**(1), 1–5. ISSN 1548-7660. URL http://www.jstatsoft.org/v20/i01.

Douglas JA (1997). "Joint Consistency of Nonparametric Item Characteristic Curve and Ability Estimation." *Psychometrika*, **62**(1), 7–28.

Douglas JA (2001). "Asymptotic Identifiability of Nonparametric Item Response Models." *Psychometrika*, **66**(4), 531–540.

Douglas JA, Cohen A (2001). "Nonparametric Item Response Function Estimation for Assessing Parametric Model Fit." *Applied Psychological Measurement*, **25**(3), 234–243.

Eubank RL (1988). *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, New York.

Härdle W (1990). *Smoothing Techniques with Implementation in S*. Springer, New York.

Härdle W (1992). *Applied Nonparametric Regression*, volume 19 of *Econometric Society Monographs*. Cambridge University Press, Cambridge.

Junker BW, Sijtsma K (2001). "Nonparametric Item Response Theory in Action: An Overview of the Special Issue." *Applied Psychological Measurement*, **25**(3), 211–220.

Kutylowski A (1997). "Nonparametric Latent Factor Analysis of Occupational Inventory Data." In J Rost, R Langeheine (eds.), *Applications of Latent Trait and Latent Class Models in the Social Sciences*, pp. 253–266. Vaxmann, New York.

Lei PW, Dunbar SB, Kolen MJ (2004). "A Comparison of Parametric and Non-Parametric Approaches to Item Analysis for Multiple-Choice Tests." *Educational and Psychological Measurement*, **64**(3), 1–23.

Lindsey J (1973). *Inferences from Sociological Survey Data: A Unified Approach*. Elsevier Scientific Publishing Company, Amsterdam.

Lord FM (1980). *Application of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum, Hillsdale.

Marron J, Nolan D (1988). "Canonical Kernels for Density Estimation." *Statistics & Probability Letters*, **7**(3), 195–199.

Nadaraya EA (1964). "On Estimating Regression." *Theory of Probability and Its Applications*, **9**(1), 141–142.

Nering ML, Ostini R (2010). *Handbook of Polytomous Item Response Theory Models*. Taylor

& Francis Group, New York.

Olsson U, Drasgow F, Dorans N (1982). "The Polyserial Correlation Coefficient." *Psychometrika*, **47**(3), 337–347.

Ostini R, Nering ML (2006). *Polytomous Item Response Theory Models*. SAGE Publications, London.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ramsay JO (1991). "Kernel Smoothing Approaches to Nonparametric Item Characteristic Curve Estimation." *Psychometrika*, **56**(4), 611–630.

Ramsay JO (1997). "A Functional Approach to Modeling Test Data." In WJ van der Linden, RK Hambleton (eds.), *Handbook of Modern Item Response Theory*, pp. 381–394. Springer, New York.

Ramsay JO (2000). *TestGraf: A Program for the Graphical Analysis of Multiple Choice Test and Questionnaire Data*. URL http://www.psych.mcgill.ca/faculty/ramsay/ramsay.html.

Ramsay JO, Abrahamowicz M (1989). "Binomial Regression with Monotone Splines: A Psychometric Application." *Journal of the American Statistical Association*, **84**(408), 906–915.

Rice J (1984). "Bandwidth Choice for Nonparametric Regression." *The Annals of Statistics*, **12**(4), 1215–1230.

Samejima F (1981). "Efficient Methods of Estimating the Operating Characteristics of Item Response Categories and Challenge to a New Model for the Multiple-Choice Item." *Technical report*, University of Tennessee, Department of Psychology.

Silverman BW (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

Simonoff JS (1996). *Smoothing Methods in Statistics*. Springer, New York.

The Voluntary HIV-1 Counseling and Testing Efficacy Study Group (2000a). "Efficacy of Voluntary HIV-1 Counseling and Testing in Individual and Couples in Kenya, Tanzania, and Trinidad: A Randomised Trial." *Lancet*, **356**(9224), 103–112.

The Voluntary HIV-1 Counseling and Testing Efficacy Study Group (2000b). "The Voluntary HIV-1 Counseling and Testing Efficacy Study: Design and Methods." *AIDS and Behavior*, **4**(1), 5–14.

Thissen D, Steinberg L (1986). "A Taxonomy of Item Response Models." *Psychometrika*, **51**(4), 567–577.

van der Linden WJ, Hambleton RK (1997). *Handbook of Modern Item Response Theory*. Springer, New York.

Watson GS (1964). "Smooth Regression Analysis." *Sankhy Aitcl̂ose, Series A*, **26**(4), 359–372.

Wickelmaier F, Strobl C, Zeileis A (2012). "Psychoco: Psychometric Computing in R." *Journal of Statistical Software*, **48**(1), 1–5. ISSN 1548-7660. URL http://www.jstatsoft.org/v48/i01.

Wong WH (1983). "On the Consistency of Cross-Validation in Kernel Nonparametric Regres-

sion." *The Annals of Statistics*, **11**(4), 1136–1141.

Zumbo BD (2007). "Three Generations of Differential Item Functioning (DIF) Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going." *Language Assessment Quarterly*, **4**(2), 223–233.

Affiliation:

Antonio Punzo
Dipartimento di Economia e Impresa
Università di Catania
95129 Catania, Italia
Telephone: +39/095/7537640
Fax: +39/095/7537610
E-mail: antonio.punzo@unict.it
URL: http://www.economia.unict.it/punzo